Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition

Ivo Everts, Jan C. van Gemert, and Theo Gevers, *Member, IEEE*

***Abstract*— This paper considers the recognition of realistic human actions in videos based on spatio-temporal interest points (STIPs). Existing STIP-based action recognition approaches operate on intensity representations of the image data. Because of this, these approaches are sensitive to disturbing photometric phenomena, such as shadows and highlights. In addition, valuable information is neglected by discarding chromaticity from the photometric representation. These issues are addressed by color STIPs. Color STIPs are multichannel reformulations of STIP detectors and descriptors, for which we consider a number of chromatic and invariant representations derived from the opponent color space. Color STIPs are shown to outperform their intensity-based counterparts on the challenging UCF sports, UCF11 and UCF50 action recognition benchmarks by more than 5% on average, where most of the gain is due to the multichannel descriptors. In addition, the results show that color STIPs are currently the single best low-level feature choice for STIP-based approaches to human action recognition.**

***Index Terms*— Color, human activity recognition, evaluation.**

I. Introduction

HUMAN activities play a central role in video data that is abundantly available in archives and on the internet. Information about the presence of human activities is therefore valuable for video indexing, retrieval and security applications. However, these applications demand recognition systems to operate in unconstrained scenarios. For this reason, research has shifted from recognizing simple human actions under controlled conditions to more complex activities and events 'in the wild' [10]. This requires the methods to be robust against disturbing effects of illumination, occlusion, viewpoint, camera motion, compression and frame rates.

High-level approaches for unconstrained human activity recognition aim at modeling image sequences based on the detection of high level concepts [13], and may build on low-level building blocks [20] which typically consider generic video representations based on local photometric features [7], [9], [26]. High-level approaches are based on complex, computationally expensive video processing operations but may be superior to low-level approaches in terms of recognition rates. However, high-level approaches are sensitive to local geometric disturbances such as occlusion, which limits their applicability [13]. Low-level approaches are conceptually simple, relatively easy to implement and potentially sparse and efficient. Due to the local nature of features on which low-level approaches are based, they are inherently robust against recording disturbances such as occlusion and clutter. Therefore, in this paper, we focus on low-level representations for recognizing human actions in video.

Low-level action recognition approaches are often based on spatio-temporal interest points (STIPs). Here, image sequences are represented by descriptors that are extracted locally around STIP detections, see Fig. 1 for example detections. The descriptors are vector quantized based on a visual vocabulary, and subsequent learning and recognition operates on these quantized descriptors, comprising the well known bag-of-(spatio-temporal)-features framework. The formulations of spatio-temporal feature detectors and descriptors available in literature are based on single-channel intensity representations of the video data. Due to the lack of photometric invariance of the intensity channel [21], current approaches are consequently sensitive to disturbing illumination con-

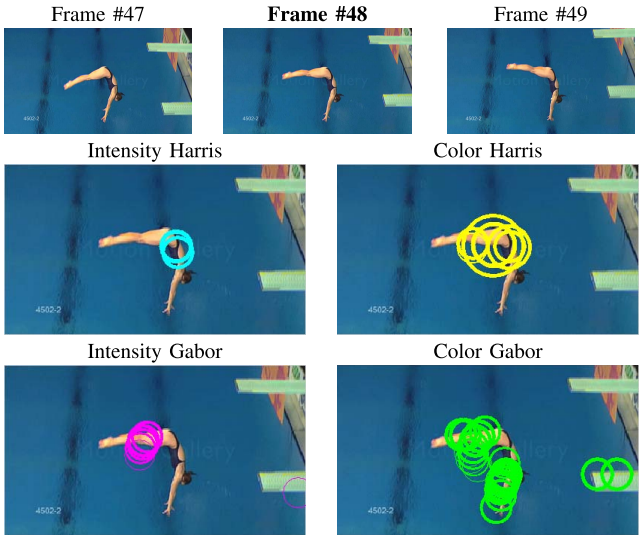

Fig. 1. Examples of STIP detections in the sequence depicted above. For illustration purposes we have polled the detectors for the 55 strongest STIPs in the original 55-frame sequence, and show the detections on frame 48 (Color online).

Manuscript received July 8, 2013; revised November 23, 2013; accepted January 8, 2014. Date of publication January 27, 2014; date of current version February 25, 2014.The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu.

I. Everts and J. C. van Gemert are with the Faculty of Science, University of Amsterdam, Amsterdam 1098, The Netherlands (e-mail: i.everts@uva.nl; j.c.vangemert@uva.nl).

T. Gevers is with the Faculty of Science, University of Amsterdam, Amsterdam 1098, The Netherlands, and also with the Computer Vision Center, Barcelona 08193, Spain (e-mail: th.gevers@uva.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2302677

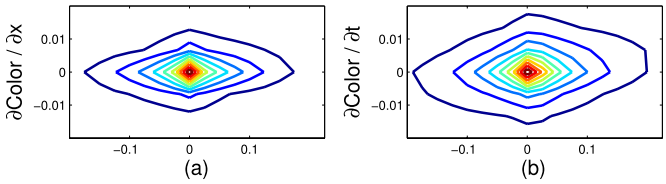

Fig. 2. Joint distributions of partial intensity and color (hue) derivatives for the spatial (a) and temporal (b) domain. The distributions are estimated from 5M pixels of one sequence of the FeEval dataset [19].

ditions such as shadows and highlights. More importantly, discriminative information is ignored by discarding chromaticity from the representation.

In the spatial (non-temporal) domain, color descriptors outperform intensity descriptors in a variety of image matching and object recognition tasks [2], [21]. The reason for this improved balance between photometric invariance and discriminative power is illustrated in Fig. 2(a) by an estimate of the joint distribution of spatial intensity and color partial derivatives, being the image features based on which descriptors are formed. The figure shows that every intensity derivative is associated with a distribution over color derivatives and vice versa. Thus, information is lost when either intensity or chromatic representations are considered in isolation. For effective feature detection and extraction based on multi-channel differential representations in the spatio-temporal domain, it is thus a precondition that similar conclusions hold for the joint distribution of temporal intensity and color derivatives. This is verified by observing Fig. 2(b), in which the joint distribution of temporal color and intensity derivatives is shown to strongly resemble the distribution of spatial derivatives in Fig. 2(a).

In this paper, we propose to incorporate chromatic representations in the spatio-temporal domain. The aim is to reformulate STIP detection and description for multi-channel video representations. Videos are represented in a variety of color spaces exhibiting different levels of photometric invariance. By this enhanced appearance modeling, we aim to increase the quality (robustness and discriminative power) of STIP detectors and descriptors for recognizing human activities in video. This is validated through a set of repeatability and recognition experiments on challenging video benchmarks. A previous version of this work appeared in [6].

A. Related Work

In the spatial domain, multi-channel photometric invariant formulations of feature detectors are reported in e.g. [18], [22], and [23]. These articles report increased repeatability, entropy, and object categorization results as compared to intensity-based detections. For descriptors, multi-channel formulations [2], [21] propose various color SIFT variants. Most notably, OpponentSIFT considerably improves the performance. Based on this, we formulate a family of increasingly invariant photometric representations which are incorporated in multi-channel formulations of spatio-temporal feature detectors and descriptors.

1) Spatio-Temporal Detectors: In the spatio-temporal domain, pioneering work by Laptev [8] extends the Harris function to 3D. Alternatively, the Gabor STIP detector proposed by Dollàr et al. [4] applies a Gabor filter along the temporal axis and is not based on differential image structure. The authors [4] argue that differential based STIP detectors are incapable of detecting subtle and periodic motion patterns. Gabor STIPs are therefore essentially different from Harris STIPs and we develop multi-channel formulations for both detectors to study differential as well as raw spatio-temporal image data.

As an alternative to STIP-based sampling, local descriptors may be extracted along motion trajectories [25]. Here, densely sampled points are tracked from frame to frame based on optical flow. As the method involves tracking and dense multi-scale optical flow computation, the associated computational complexity is typically higher than that of STIP-based approaches. Depending on the descriptor(s) that are subsequently extracted, this sampling method may compare favorably in terms of recognition rates. In this paper, we focus on the sparser STIP-based approach for studying color in the spatio-temporal domain.

Other color STIPs have been proposed earlier in [17]. However the formulation of the multi-channel spatio-temporal structure tensor for the 3D Harris function is somewhat erroneous. Also, the proposed color STIP descriptor is a concatenation of a color histogram, an intensity-based gradient (HOG) and optical flow (HOF) descriptor, which is not shown to produce performance improvements with respect to other existing STIP-based recognition methods. In this paper, we extend the multi-channel structure tensor of [23] in a principled manner to the spatio-temporal domain and investigate various methods to incorporate color gradients in the HOG3D descriptor.

2) Spatio-Temporal Descriptors: Among the local spatio-temporal descriptors available in literature, the HOG3D descriptor [7] appears well-suited for large scale video representation and multi-channel extensions. In contrast to e.g. HOG/HOF [9], MoSIFT [3] or MBH [25] descriptors, the HOG3D algorithm serves as an integrated and efficient approach, as it excludes optical flow which is computationally expensive [11], [15]. Also, good results in a STIP-based bag-of-features recognition framework using the HOG3D descriptor have been achieved, especially in combination with the Gabor STIP detector [26]. Moreover, motion-based descriptors are shown in [11] to suffer from scalability issues. Therefore, we derive several multi-channel variants of the HOG3D descriptor and evaluate their performance for realistic human action recognition.

Discriminability issues associated to motion descriptors in large scale action recognition are shown in [11] to be addressed by the motion boundary histograms (MBH) of [24]. As opposed to a direct motion description, MBH is based on differential optical flow, which greatly reduces the confusion between action categories. In recent work by Wang et. al. [25], MBH descriptors extracted along motion trajectories and modeled in a multiple kernel learning framework have achieved state-of-the-art results on a large number of datasets.

Another recently proposed video descriptor for human action recognition is Gist3D [16]. This is a global descriptor

based on a 3D filter bank and describes the spatio-temporal 'gist' of a video. Reasonable recognition performance is achieved in combination with STIPs.

The works mentioned comprise low/medium level approaches to action recognition. Higher level approaches such as Action Bank by Sadanand et al. [13] give good results on some datasets. However, such high-level approaches are typically not scalable. In contrast, low-level approaches are widely applicable, conceptual simple, sparse and exhibit reasonable computational complexity. Moreover, they may serve as powerful building blocks for higher level methods [20]. We contribute by considering a variety of photometric representations for STIP detection and description for enhancing low-level approaches to action recognition.

II. Photometric Representations

We model the formation of images by the dichromatic reflection model [14],

$$\mathbf{f} = e(m^b \mathbf{c}^b + m^i \mathbf{c}^i), \tag{1}$$

where $\mathbf{f} = (R, G, B)^T$ is the sum of the body reflectance color $\mathbf{c}^b$ with the interface reflection color $\mathbf{c}^i$. The contributions of these reflectance colors are weighted by their respective magnitudes m^b and m^i, that depend on the surface orientation and illumination direction. Additionally, the specular reflection m^i is viewpoint dependent. The intensity of the light source is represented by e.

Invariance against highlights (shifts in the signal) can be achieved by representations that cancel out the additive interface reflection term $m^i \mathbf{c}^i$. Signal scalings, such as those caused by shadows and shading, are ignored by dividing out the light source intensity e. Here, we consider the transformation of the RGB image to the opponent color space [2], [5], [21], [22]

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} R - G \\ R + G - 2B \\ R + G + B \end{pmatrix}. \tag{2}$$

The transformation approximately decorrelates the image channels, resulting in intensity O_3 and chromatic components O_1, O_2. Based on these formulations, several photometric properties can be derived.

Highlights. Due to subtraction of RGB components in eq. (2), the reflection term from eq. (1) is subtracted in the formulations of O_1 and O_2. Hence, the chromatic opponent components are invariant to signal shifts such as those caused by (white) highlights.

Shadow-shading. The chromatic components are normalized by intensity O_3, canceling out the light source intensity term from eq. (1). This yields the shadow and shading invariants $\left[\frac{O_1}{O_3}, \frac{O_2}{O_3}\right]$.

Shadow-shading-highlights. Invariance against both scalings and shifts in the signal is achieved by considering the ratio of chromatic components: $\frac{O_1}{O_2}$. This results in the shadow-shading-highlight invariant hue representation.

We refer to these photometric image representations as I(intensity), C(hromatic), N(ormalized chromatic) and H(ue). These can be ordered with respect to their invariance level: $H \succ N \succ C \succ I$. The intensity I preserves most image structures, which is the most discriminative representation. Therefore the intensity-normalized representations N and H have a higher level of photometric invariance than C, in which the light source intensity is preserved. We summarize the representations and their properties in Table I.

TABLE I
PHOTOMETRIC IMAGE REPRESENTATIONS. CHROMATIC COMBINATIONS WITH THE INTENSITY CHANNEL YIELD *IC*, *IN* AND *IH*

	Intensity	Chromatic	N-Chromatic	Hue
Representation	O_3	$[O_1, O_2]$	$\left[\frac{O_1}{O_3}, \frac{O_2}{O_3}\right]$	$\frac{O_1}{O_2}$
Invariant to	-	Highlights	Shadows	Hl. & Sh.
Reference	I	C	N	H

The lack of discriminative power associated with the chromatic representations C, N and H typically renders them unsuitable for matching and recognition tasks. Combinations of intensity and chromatic channels result in IC, IN and IH. Note that the three-channel representation IC comprises the original opponent channels $[O_1, O_2, O_3]$. These representations are established first, i.e. , prior to any subsequent processing. All channels are min-max normalized using the theoretical extremal values per channel based on the transformations in eq. (2) and Table I so as to weight them equally a-priori.

III. Multi-Channel STIP Detection

Multi-channel Harris STIPs. Harris STIPs are local maxima of the 3D Harris energy function based on the structure tensor [8]. A multi-channel formulation of the structure tensor has been developed in e.g. [23] which prevents opposing color gradient directions to cancel each other out. Here, we incorporate multiple channels in the spatio-temporal structure tensor [8].

The multi-channel volume $\mathbf{V}$ consisting of n_c channels is denoted by $\mathbf{V} = (V^1, V^2, ..., V^{n_c})^T$. The individual channels are represented in scale space $V^j = g(\cdot; \sigma_o, \tau_o) * f^j(\cdot)$, where $g(\cdot; \cdot, \cdot)$ is the 3D Gaussian kernel with equal scales along the spatial dimensions, σ_o and τ_o are the spatial and temporal observation scales and $f^j : \mathbb{R}^3 \to \mathbb{R}$ is the imaging function of channel j. The multi-channel spatio-temporal structure tensor is then defined by

$$\mathbf{S} = g(\cdot; \sigma_i, \tau_i) * \begin{pmatrix} \mathbf{V}_x \cdot \mathbf{V}_x & \mathbf{V}_x \cdot \mathbf{V}_y & \mathbf{V}_x \cdot \mathbf{V}_t \\ \mathbf{V}_y \cdot \mathbf{V}_x & \mathbf{V}_y \cdot \mathbf{V}_y & \mathbf{V}_y \cdot \mathbf{V}_t \\ \mathbf{V}_t \cdot \mathbf{V}_x & \mathbf{V}_t \cdot \mathbf{V}_y & \mathbf{V}_t \cdot \mathbf{V}_t \end{pmatrix}, \tag{3}$$

where σ_i and τ_i denote the spatial and temporal integration scale respectively. In Fig. 3 we illustrate the response per representation. Incorporating increasingly invariant photometric representations clearly has an effect on the Harris energy. The highlight on the shiny heart-shaped object surface part triggers a strong response for the original I-based energy functions. This effect is clearly dampened in the C representation. However, the reflected illumination by the colored matte-shiny (left) object part still triggers a response, as the nature of the local object surface causes signal changes that are not captured

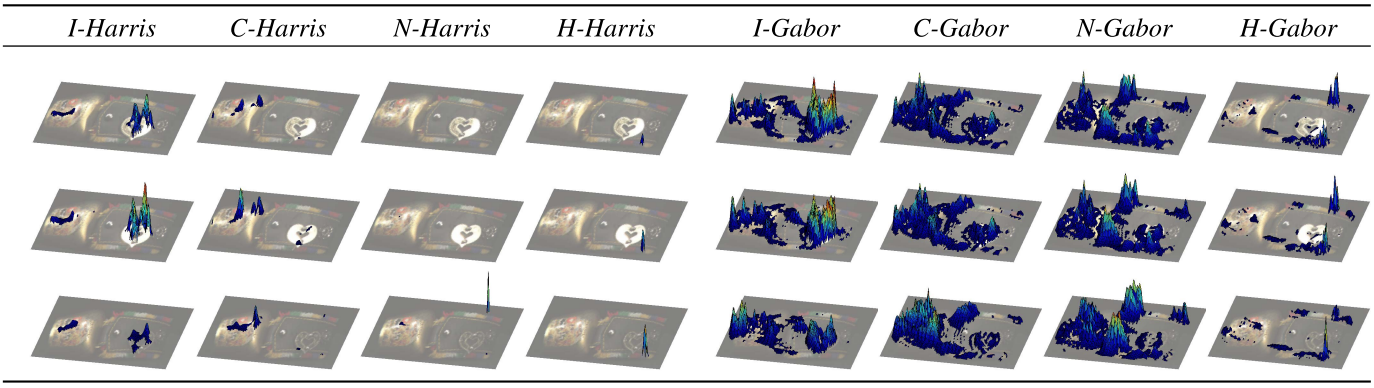

Fig. 3. Superimposed Harris and Gabor responses for **I**ntensity, **C**hromatic, **N**ormalized chromatic and **H**ue on three images of a rotating object on which a strong highlight is present. The Harris energy function mainly responds to differential changes in the signal, whereas the Gabor function fires on general spatio-temporal fluctuations. Note the dampened response to the highlight in the invariant channels.

by a simple shift. Intensity normalization of the chromatic components (N) then causes this response to be dampened, while emphasizing colorful transitions on the object surface. Finally, the scaling- and shift- invariant H representation eliminates essentially all responses except for salient color transitions.

Multi-channel Gabor STIPs. The Gabor STIP detector is based on a Gabor filtering procedure along the temporal axis [4]. Invoking multiple channels is straightforward because the energy function is positive by formulation. Hence, no additional care has to be taken to account for conflicting response signs between channels

$$R = \sum_{j=1}^{nc} (g(\cdot;\sigma_o) * h_{ev}(\cdot;\tau_o) * V^j)^2 + (g(\cdot;\sigma_o) * h_{od}(\cdot;\tau_o) * V^j)^2. \quad (4)$$

Here, the 2D Gaussian smoothing kernel $g(\cdot;\cdot)$ is applied spatially, whereas the Gabor filter pair $\{h_{ev}(\cdot;\cdot), h_{od}(\cdot;\cdot)\}$ measures the periodicity of the observed signal along the temporal dimension. As illustrated in Fig. 3, the I-Gabor energy is mainly clustered around an incidental highlight, whereas the response-triggering local photometric events become increasingly rare and colorful along with the level of photometric invariance level of the representation.

IV. Multi-Channel STIP Description

The HOG3D descriptor [7] is formulated as a discretized approximation of the full range of continuous directions of the 3D gradient in the video volume. That is, the unit sphere centered at the gradient location is approximated by a regular n-sided polyhedron with congruent faces. Tracing the gradient vector along its direction up to intersection with any of the polyhedron faces identifies the dominant quantized direction. Quantization proceeds by projecting the gradient vector on the axes running through the gradient location and the face centers with a matrix multiplication of the 3D gradient vector $\boldsymbol{g}$,

$$\boldsymbol{q} = (q_1, \ldots, q_n)^T = \frac{P \cdot \boldsymbol{g}}{||\boldsymbol{g}||_2}, \quad (5)$$

where P is the $n \times 3$ matrix holding the face center locations and $\boldsymbol{q}$ is the projection result (i.e. the histogram of 3D gradient directions). Note that the contribution is distributed among nearby polyhedron faces. Descriptor dimensionality may be reduced by allocating opposing gradient directions to the same orientation bin. The descriptor algorithm proceeds by centering a cuboid at the STIP location, which is tessellated into a spatio-temporal grid. Histograms are computed for every grid cell and concatenated to form the final descriptor [7].

Chromaticity is incorporated in the HOG3D descriptor by considering the representations from section (II) in a multi-channel formulation of the gradient vector $\boldsymbol{g}$ in eq. (5). We follow the standard practice of concatenation of the per-channel descriptors [2], [5], [21]:

$$\boldsymbol{g'} = \{\boldsymbol{g}^j\},\ j = 1, \ldots, n_c. \quad (6)$$

We also compute a single gradient variant where we prevent the effect of opposing color gradient directions by using tensor formulations. In tensors, opposing directions reinforce each other by summing the gradient *orientations* as opposed to their *directions* [23],

$$\boldsymbol{g''} = \sum_{j=1}^{nc} \boldsymbol{g}^j \cdot \boldsymbol{g}^j. \quad (7)$$

This formulation of the gradient defines half of the full sphere of directions which is one of the HOG3D flavors in [7]. Here, it naturally follows from a tensor formulation of the multi-channel 3D gradient.

We formulate another variation as the summation of per-channel full direction descriptors. Together with the tensor-based approach, we call this descriptor *integration* as opposed to *concatenation*. The variant benefits from the expressiveness associated with the full set of multi-channel directions while maintaining the same dimensionality as a single channel descriptor. Note that the differences between integration and concatenation of channels do not apply to single-channel descriptors. The descriptor variants and their associated dimensionalities are summarized in Table II.

TABLE II
MULTI-CHANNEL HOG3D VARIANTS. $\mathfrak{C}$ DENOTES SOME PHOTOMETRIC REPRESENTATION COMPRISING n_c CHANNELS. THE DIMENSIONALITY OF AN INTEGRATED DIRECTION-BASED DESCRIPTOR IS CONSIDERED DEFAULT ($1D$, WHICH IS 360 IN THIS PAPER), BASED ON WHICH WE DERIVE THE DIMENSIONALITY OF THE OTHER DESCRIPTOR VARIANTS. VARIANTS OF $\mathfrak{C}$ ARE DENOTED BY SUBSCRIPT FLAGS, INDICATING CHANNEL COMBINATION (INTEGRATION/CONCATENATION) AND GRADIENT QUANTIZATION (ORIENTATION/DIRECTION)

	Gradient Orientation	Gradient Direction
Channel Integration	$\mathfrak{C}_{1,1} : D/2$	$\mathfrak{C}_{1,0} : 1D$
Channel Concatenation	$\mathfrak{C}_{0,1} : n_cD/2$	$\mathfrak{C}_{0,0} : n_cD$

Fig. 4. Examples from FeEval dataset. From left to right: original, noise, darken.

Fig. 5. Examples from UCF sports, UCF11 and UCF50 datasets (images are cropped).

V. EXPERIMENT

We evaluate the multi-channel STIP detectors and descriptors through a set of repeatability and action recognition datasets.

A. Implementation Details and Notation

We base our implementation of STIP detectors on the activity recognition toolbox by Dollàr et al. [4] while re-implementing the HOG3D descriptor of Kläser et al. [7].

STIP scale: For the Gabor detector, we set the spatial scale $\sigma_o = 2$ and the temporal scale $\tau_o = \sqrt{8}$ in eq. (4). Note that this setting for τ_o is in conflict with e.g. [26], but we have found that the proposed default setting of $\tau_o = 4$ is too large for descriptor extraction in short sequences. For the Harris detector, we consider a reduced set of spatial scales with respect to prior work, as we have found this to be satisfactory in terms of discriminative power and computational load. Specifically, for computing the Harris energy based on eq. (3), we consider $\sigma_o = \sqrt{2^i}, i \in \{2, 3, 4\}$ and $\tau_o = \sqrt{2^j}, j \in \{1, 2\}$. As in e.g. [26] and [9], we do not perform STIP scale selection because of its high computational costs and decreased recognition performance [8].

Cuboids: Descriptors are extracted from cuboids centered at STIP locations. The spatio-temporal extent as well as the grid layout of these cuboids may be discriminatively optimized such as in [7]. In this paper, we refrain from such an optimization scheme in order to maintain focus on the integration of chromatic channels. Instead, we consider one particular setting (from e.g. [26]) in which the extent of a cuboid is defined as $\Delta_x = \Delta_y = 18\sigma_o$ and $\Delta_t = 8\tau_o$. For feature aggregation, we employ a $3 \times 3 \times 2$ spatio-temporal pooling scheme. This grid layout is attractive due its compactness, whereas we have not found significant dependencies of our results on these settings for our purpose.

Descriptors: We consider the four variants of the multi-channel HOG3D descriptor as summarized in Table II. The variants are denoted by flagging the descriptor names. The first flag denotes whether the descriptor channels are integrated (or otherwise concatenated), whereas the second flag denotes the usage of gradient orientations (as opposed to directions). For example, $IC_{0,1}$ denotes the concatenated orientation-based Opponent-HOG3D descriptor. Integrated, orientation-based descriptors such as $IN_{1,1}$ follow from the tensor-based approach in eq. (7). There is no difference between $I_{0,\cdot}$ and $I_{1,\cdot}$ as I comprises a single channel.

We use integral video histograms for aggregating features over grid cells. We refrain from gradient approximation based on integral video representations of the partial derivatives as in [7], because this affects the information that we wish to study. For descriptor normalization, we adopt the method proposed by Brown et al. [1] in which the normalization cut-off threshold is a discriminatively optimized function of the descriptor dimensionality. By this, we discard the time consuming task of determining the optimal normalization parameters per descriptor variant.

In summary, apart from the photometric representations, our HOG3D implementation differs slightly from the original version [7] by 1) exact gradient computation, 2) descriptor normalization and 3) spatio-temporal pooling.

Recognition. Based on the multi-channel STIP detectors and descriptors, we perform action recognition in a standard bag-of-features learning framework. Unless stated otherwise, we closely follow the setup of [26]. Here, codebooks are created by clustering 200K randomly sampled HOG3D descriptors using k-means in 4000 clusters. A sequence is then represented by quantizing the extracted HOG3D descriptors based on the learned codebook. An SVM is trained based on the χ^2 distance between codebook descriptors. Evaluation of the learned classifier is usually performed in a leave-n-out cross validation setup. Every experiment is repeated three times for different codebooks, which produces typical standard deviations between 0.2 and 1 percentage point (depending on dataset size and the number of STIP detections).

B. Datasets

We measure STIP repeatability and descriptor entropy for videos taken from the **FeEval** dataset [19]. This dataset consists of 30 videos taken from television series, movies and lab recordings where each video is artificially distorted by applying different types of photometric and geometric transformations. Every transformation type is associated to a challenge, in which the distortion is applied in increasingly severe steps. We consider the videos from the television series up to the first occurring shot boundary. That is, we do not aim

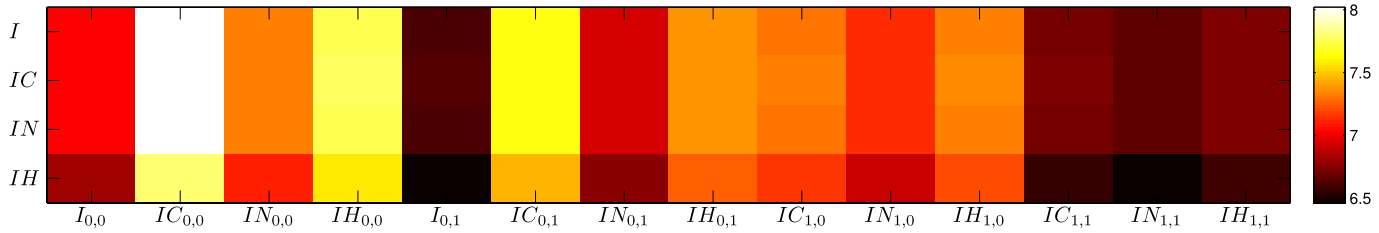

Fig. 6. Entropy of descriptor variants extracted around STIPs from several detector variants. Multi-channel descriptors are associated to higher entropies than their single-channel counterparts. This holds for both integration and concatenation of channels. The figure looks similar for Harris and Gabor STIPs.

TABLE III
STIP REPEATABILITY FOR MULTI-CHANNEL HARRIS AND GABOR DETECTORS BASED ON THE CONSIDERED PHOTOMETRIC REPRESENTATIONS

	I	IC	IN	IH	C	N	H
Harris	**61.3%**	**61.6%**	**61.3%**	37.0%	45.6%	40.5%	28.7%
Gabor	43.6%	43.6%	43.6%	24.4%	25.4%	22.9%	19.3%

at studying STIP behavior in controlled settings, cartoons or in typical movie settings for which editing effects are frequent. We consider the full set of challenges: blur, compression, darken, lighten, median filter, noise, sampling rate and scaling and rotation. Some examples are shown in Fig. 4.

For an in-depth evaluation of detector and descriptor settings we use the **UCF sports** dataset [12]. The dataset exhibits 10 sports action categories in 150 videos, all of which are horizontally flipped to increase the dataset size. Performance is evaluated in a leave-one-out cross validation scheme, in which the flipped version of the considered test video is removed from the training set. The best performing experimental settings are applied to the **UCF11** [10] and **UCF50** [11] datasets. The datasets contain 11 and 50 human action classes in about 1200 and 6700 videos respectively; UCF50 is a superset of UCF11. These challenging datasets comprise youtube videos exhibiting real human activities. Here, performance is evaluated through a leave-one-group-out cross validation scheme over 25 groups, in which we exactly follow the authors' guidelines.[1] See Fig. 5 for some examples of the datasets.

C. STIP Repeatability

We poll the detectors for an average number of 10 STIPs per frame of the FeEval videos. A repeatability score is obtained by considering the detections in the challenge sequence, and computing the relative overlap of the cuboid around the detected STIP location with the corresponding location in the original sequence. We take the spatio-temporal extent of the cuboid to be equal to the observation scale. The repeatability scores averaged over all sequences and challenges are presented in Table III.

Harris STIPs are more stable than Gabor STIPs. Nonlinear differential spatio-temporal signal changes are more distinctive than temporal fluctuations only. As the representation becomes increasingly invariant, repeatability progressively decreases. Also, combining the invariants with intensity does not increase repeatability with respect to using intensity only (marginal improvements for the IC representation). Moreover, the IH representation attains lower repeatability scores than I. The reason for these reduces scores is that, as disturbing conditions are effectively ignored, so are spatio-temporal image structures on which stable STIPs are detected. Adding C or N to the intensity I basically leaves the repeatabililty unaltered for this dataset. However, the STIP discriminability experiments will show different recognition scores for these representations.

From here on, the pure chromatic representations are discarded from the experimental batch due to the associated lack of discriminative power.

D. Descriptor Entropy

Here, we study the amount of information contained in each of the considered descriptors. For this, we extract unnormalized descriptors from the cuboids around STIP detections in the set of undistorted FeEval videos. The descriptors D_i are then L_1-normalized to allow for the computation of entropy:

$$entropy(D_i) = -\sum_{j=1}^{|D_i|} D_i^j log_2(D_i^j). \quad (8)$$

The above is illustrated in Fig. 6 for Gabor STIPs. Entropies are averaged over all descriptors and sequences. The figure is essentially similar for descriptors extracted around Harris STIPs.

Standing out from the figure is the high entropy associated to the $IC_{0,0}$ descriptor (i.e. concatenated direction-based Opponent-HOG3D). This is partly explained by its high dimensionality due to concatenation. Note however the increased entropy with respect to $IN_{0,0}$, which has the same dimensionality. In that respect it also stands out that the entropy associated to the 2-channel descriptor $IH_{0,0}$ is higher than that of the 3-channel descriptor $IN_{0,0}$. We conclude from this that the chromatic ratio constituting H exhibits more (differential) variation than the intensity-normalized channels in N, whereas most variance is associated with C.

The single-channel descriptor $I_{0,0}$ is associated with a considerable lower entropy than its multi-channel counterparts. These differences are dampened when the channels are integrated instead of concatenated, by which the multi-channel dimensionality is equalized to that of a single channel.

[1] http://crcv.ucf.edu/data/UCF50.php

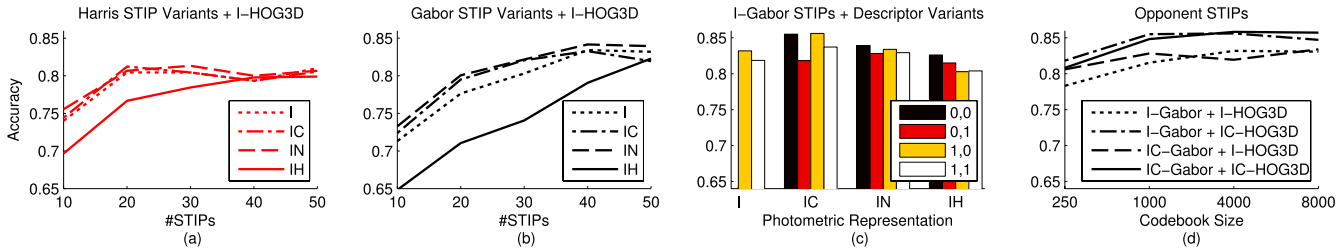

Fig. 7. Recognition performance on the UCF sports dataset per photometric representation for varying amounts of Harris (a) and Gabor (b) STIPs. Influence of the photometric representations on descriptor variants (c). Combinations of the top-performing IC-Gabor STIPs and $IC_{1,0}$-descriptors for varying codebook sizes (d).

However, the integrated descriptors $IC_{1,0}$ and $IH_{1,0}$ are still clearly associated to higher entropies, whereas the difference between $IN_{1,0}$ and $I_{0,0}$ is marginal.

Orientation-based descriptors exhibit lower entropies than direction-based descriptors. This follows from their definition: two opposing gradient directions are indistinguishable in terms of their orientation. Observations regarding photometric representations and channel integration with respect to the direction-based descriptors also hold for orientation-based descriptors.

With respect to varying photometric representations in the detector, we observe a considerable drop in entropy for the IH detector as compared to the other representations. This is explained by the fact that H causes the detector to fire on signal fluctuations that do not necessarily correspond to strong structures in the intensity profile. There appears no substantial differences between the other representations, although slightly higher entropies are attained for IC detections.

E. Color STIP Detector Discriminability

For evaluating action recognition performance on the UCF sports dataset, we consider the photometric variants of both the Harris and Gabor detectors. Direction-based intensity HOG3D ($I_{,0}$) descriptors are extracted around multi-channel STIP detections, so as to seperate the analyses regarding STIP detection and description. Recognition accuracy is computed for an average of $\{10, 20, 30, 40, 50\}$ STIPs per frame by varying the detection threshold. Results are given in Fig. 7(a) and (b).

We first validate our implementation by comparing recognition accuracies with the evaluation reported on intensity in [26]. Here, the average number of Harris STIPs is 33, for which an accuracy of 79.9% is attained. We obtain 80.4% for 30 STIPs per frame. As for the Gabor detector, [26] reports an accuracy of 82.9% for 44 STIPs. This is comparable to our performance of 83.4% for 40 STIPs.

1) Color STIPs: It is shown in Fig. 7(a) and (b) that discriminative power is severely hampered by integrating H in the energy functions. This is expected because H is associated to the highest level of photometric invariance. As more detections are requested, however, performance converges to that of I-STIPs. Considering Harris STIPs in Fig. 7(a), integrating the C and N representations leads to marginal performance differences compared to I. For small to moderate amounts of STIPs, recognition accuracy is somewhat improved, in particular for IN. The primary characterization of Harris STIPs in terms of distinctiveness and sparsity is mainly due to nonlinear fluctuations in the spatio-temporal intensity signal. Adding chromatic components to the formulation of the energy function does not drastically alter this characterization.

Regarding the multi-channel Gabor detector in Fig. 7(b), discriminative STIPs are detected for the C and especially N channels as compared to using I alone. While I by itself contains the most important information regarding spatio-temporal signal fluctuations, invariants may prevent the detector to fire on disturbing factors such as highlights and shadows. Also, we assume the specific colorfulness of local spatio-temporal events associated to certain actions to be informative (e.g. 'Diving' (skin color, blue water) and 'Riding-Horse' (brown horse, green field and trees)).

2) Discussion on Sparsity, Distinctiveness and Scale: Harris STIPs are more discriminative than Gabor STIPs for a relatively small number of detections. This relative performance difference reverses as more STIPs are considered. The reason for this is related to sparsity, distinctiveness and scale.

As can be derived from Fig. 3, the Harris function is sparser than the Gabor energy. The Harris function fires only on relatively rare events - nonlinear signal changes in both space and time - which are also distinctive in scale space, and are usually caused by human activity rather than background and/or camera motion. As a consequence, Harris STIPs are highly discriminative, but very sparse: there resides a large and indifferent gap between the thresholds of a good quality Harris STIP detector and a noise detector. Opposed to this, the Gabor detector is more generic and covers the image sequences more densely. This results in improved recognition results as more STIPs are requested, whereas the performance of the Harris detector as a function of the number of STIPs quickly plateaus and even degrades.

Whereas the Harris function is typically computed over multiple scales, the Gabor detector (as originally proposed) operates at a single scale. In fact, we have found in the recognition experiments in which we poll the detectors for a fixed number of interest points, that a multi-scale Gabor implementation seriously hampers the recognition performance (results not shown). The reason for this is that the across-scale Gabor responses are highly correlated. This results in overly redundant overlapping detections for local volumes exhibiting strong periodic signal fluctuations, whereas other discriminative local

TABLE IV

COLOR STIP ACTION RECOGNITION RESULTS ON UCF11 AND UCF50 DATASETS. THE FIRST 5 COLUMNS SHOW RESULTS FOR DIRECTION-BASED DESCRIPTORS, WHEREAS RESULTS FOR ORIENTATION-BASED DESCRIPTORS ARE SHOWN IN THE REMAINING COLUMNS

	Detector/Descriptor:	$I_{.,0}$	$IC_{1,0}$	$IC_{0,0}$	$IN_{1,0}$	$IN_{0,0}$	$I_{.,1}$	$IC_{1,1}$	$IC_{0,1}$	$IN_{1,1}$	$IN_{0,1}$
UCF11	$I-Gabor$	73.8%	77.5%	78.2%	76.0%	76.4%	71.6%	75.8%	74.2%	73.8%	74.6%
	$IC-Gabor$	73.8%	78.4%	78.1%	76.6%	76.3%	71.5%	75.4%	73.7%	73.9%	74.3%
	$IN-Gabor$	**74.5**%	77.5%	**78.6**%	76.7%	76.4%	72.4%	76.0%	74.6%	74.2%	74.0%
UCF50	$I-Gabor$	68.3%	71.7%	70.9%	71.2%	72.1%	**68.8**%	72.6%	69.7%	71.8%	72.0%
	$IC-Gabor$	68.5%	71.8%	70.8%	71.2%	71.9%	**68.8**%	72.4%	69.8%	71.5%	72.4%
	$IN-Gabor$	68.4%	71.8%	71.1%	71.0%	71.8%	68.5%	**72.9**%	69.9%	71.6%	72.5%

volumes may not be detected at all. Applying the Gabor filters at a single scale only is therefore not so much a choice of design; it is rather instrumental to the method. These arguments do not apply to the Harris detector due to its associated sparsity, i.e. single scale Harris STIPs are insufficient for effective recognition. The unnecessity of multi-scale processing grants a large advantage to the Gabor detector over the Harris detector in terms of computational efficiency. The experimental summary over all datasets in Table VI shows the effectiveness of the Gabor detector.

F. Color STIP Descriptor Discriminability

For the action recognition experiments on the UCF sports dataset, descriptors are extracted around Gabor STIPs as these have shown superior recognition performance over Harris STIPs in Fig. 7(a) and (b). The detector representation is fixed to I. We adopt the detection threshold that yields 50 STIPs per frame on average. Recognition accuracies are reported in Fig. 7(c).

General conclusions about photometric invariance relate to the discriminative power of the descriptors. That is, the IC-based descriptors typically outperform IN descriptors, which in turn are favored over IH. Multi-channel descriptors usually outperform the I-based descriptor. We observe a general preference for direction-based descriptors over orientation-based descriptors (Table II). This is due to the associated wider range of expressiveness. Most apparent in this respect is the IC representation, i.e. $IC_{0,0}$ improves over $IC_{0,1}$ by almost 4 percentage points, whereas $IC_{1,0}$ attains 2 percentage points more than $IC_{1,1}$. Thus, every channel exhibits discriminative power in the full range of gradient directions. It may even be the case that the (implicit) preservation of opposing gradient directions between channels is informative. Furthermore, IC-based descriptors favor channel integration over concatenation, which is not the case for IN- and IH- based descriptors. In fact, one would expect concatenation-based descriptors to perform better in general due the enhanced expressiveness associated with multiple channels and increased dimensionality. This is also the most widely adopted approach to multi-channel descriptors, e.g. [2], [21], and [5]. However, we obtain the positive side-effect of increased recognition performance against reduced descriptor dimensionality. That is, the multi-channel descriptor dimensionality remains equal to that of a single channel. Although the difference with $IC_{0,0}$ is marginal, we report a top performance of 85.6% for $IC_{1,0}$ against 1) our $I_{.,0}$ baseline of 83.4% and 2) 82.9% reported in [26]. A summary over all datasets in Table VI illustrated the power of IC.

We conduct a final experiment on the codebook size. We consider 'Opponent STIP' combinations of I and IC Gabor STIPs with $I_{.,0}$ and $IC_{1,0}$ HOG3D descriptors. We drop the orientation-based descriptors for now. Recognition results for varying codebook sizes are depicted in Fig. 7(d). We observe that the I-IC (detector-descriptor) combination performs best up to a codebook size of 4000. Top performance is marginally improved to 85.7% by the IC-IC combination for a codebook size of 8000. The computational load associated to such a vocabulary is not worth the effort, considering the performance of 85.5% attained by the I-IC combination for a much smaller codebook size of 1000. We have not observed a relationship between descriptor dimensionality and codebook size.

In contrast to these low/medium level action recognition approaches, the high level Action Bank approach of [13] reaches an accuracy of 95% on UCF sports. Here, we focus on low-level approaches, and our best performance for 50 STIPs per frame is on par with the performance of 85.6% for densely sampled I-HOG3D descriptors in [26], which on average yields over 600 descriptors per frame. Based on a combination of HOG, HOF and MBH descriptors extracted along dense motion trajectories, a performance of 88% is achieved in [25]. Compared to this, our STIP-based approach does a good job considering that it outperforms all reported individual features on UCF sports.

G. UCF11

Based on the in-depth evaluations on UCF sports, we select the I, IC and IN representations for both STIP detection and description for evaluation on the UCF11 and UCF50 datasets. Results are presented in Table IV and summarised in Table VI.

Differences between performance in the detectors are again small, but we observe a consistent top-performing combination of IN-Gabor STIPs with IC-based HOG3D. Thus, we conclude that a certain amount of invariance against local photometric events is beneficial for STIP detection, whereas the descriptor should be extracted from the most discriminative representation.

We achieve a baseline result of 73.8% on the UCF11 dataset for the intensity-based STIP variant. Adding chromaticity increases the recognition accuracies substantially. Also here, best performance is achieved by the direction-based IC descriptors: 78.4% for $IC_{1,0}$ on IC-Gabor STIPs and 78.6%

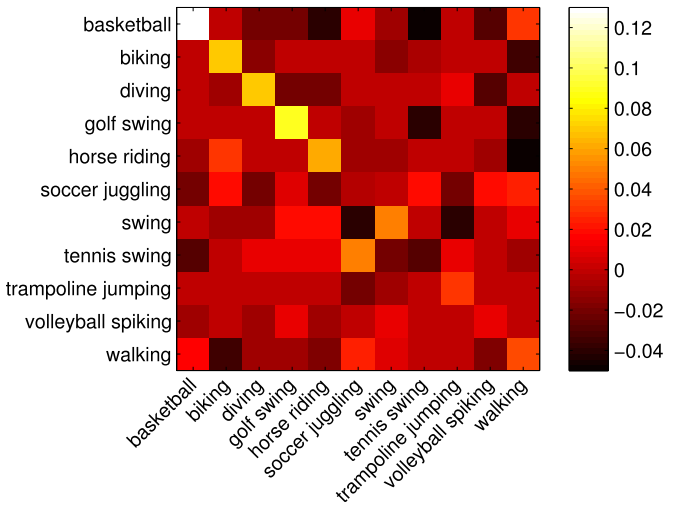

Fig. 8. Confusion difference matrix between UCF11 categories. Depicted is the element-wise difference between the confusion matrices of (best performing) color and intensity STIPs.

for $IC_{0,0}$ on IN-Gabor STIPs. The representation of the detector appears to be more influential on this dataset, although its contribution is marginal on average.

The results compare favourably to the trajectory-based harvesting of HOG and HOF features in [25], for which 72.6% and 70% is achieved respectively. However, they report a superior performance of 84.1% for their motion boundary histograms.

1) Discussion on Inter-Class Confusion: For a detailed analysis of the results on UCF11 we have included a confusion-difference matrix in Fig. 8. The usage of color causes most performance gain for the category 'basketball'. Corresponding videos in the dataset exhibit mostly practicing individuals, whereas considerable variations are observed in other facets such as indoor/outdoor, solid/shaking camera work and clothing. These observations are supportive for the argument that multi-channel processing is useful for feature extraction in general, irrespective of the actual color itself. In addition to this, category-specific motion patterns are more accurately described by using color. For example, a basketball generally has the same orange color, which makes the description of its associated motion (bouncing) more accurate. Furthermore, the usage of color decreases the confusion between 'basketball' and 'horse riding', and especially 'tennis swing'. The initial confusion (i.e. based on intensity-STIPs) between 'basketball' and 'tennis swing' is comprehensible, as most videos of both categories exhibit, in general, an individual performing the activities in isolation. Specific information associated to e.g. the colors of the basketball and tennis courts alleviate much of the confusion. The same line of reasoning applies to the confusion between 'tennis swing' and 'golf swing', and to a lesser extent 'basketball' and 'vollyball spiking', as the associated videos exhibit a single, sudden burst of activity performed by an individual. Less evident is the reason for resolved confusion between 'basketball' and 'horse riding'. Videos associated to the latter exhibit a walking or galopping horse, which is characterized by a periodic motion pattern resembling that of a person shooting a basketball. It is probably the case that a bouncing basketball also renders similar motion patterns, while its color then provides the power to discriminate. Opposed to this, it stands out that color STIPs increase the confusion between 'tennis swing' and 'soccer juggling'. This is mainly due to the fact that in one 'soccer juggling' video group the activity is performed on a typical tennis hardcourt, which renders similar patterns in all color channels.

TABLE V
RECENT UCF50 RESULTS AVAILABLE IN LITERATURE

Ref.	Description	%
[25]	Trajectory(All)	84.5%
	Trajectory(MBH)	82.2%
	Trajectory(HOF)	68.2%
	Trajectory(HOG)	68.0%
[15]	Dense(All)	83.3%
	Dense(MBH)	80.1%
	Dense(HOG3D)	72.4%
	Dense(HOF)	69.7%
	Dense(HOG)	58.6%
[11]	Scene context + STIP(MBH)	76.9%
	Scene Context	47.6%
	STIP(MBH)	71.9%
[16]	Gist3D + STIP(HOG/HOF)	73.7%
	Gist3D	65.3%
	STIP(HOG/HOF)	54.3%
[13]	Action Bank	57.9%
	STIP(HOG/HOF)	47.9%
Here	Color STIP(HOG3D)	72.9%

H. UCF50

Considering the results on UCF50 in Table IV, we observe that best performance is achieved with orientation-based descriptors, as opposed to the direction-based descriptors that are favoured for UCF sports and UCF11. As the number of categories increases, descriptor robustness becomes more important. We observe a baseline result of 68.8% for $I_{\cdot,1}$. This is substantially higher than the results reported in [13] for Action Bank (57.9%) and Harris STIP + HOG/HOF (47.9%) (see Table V for an overview of recent results on UCF50). We conclude that the Action Bank method is not scalable and suffers from increased geometric variations. As for Harris STIP + HOG/HOF, we conclude that the high degree of distinctiveness of spatio-temporal corners limits generalization capacity for these descriptors. A performance of 76.9% is reported in [11] for a combination of scene context and spatio-temporal descriptors. Here, the best performing spatio-temporal descriptor is MBH on Harris STIPs, which achieves 71.9%. This shows the generalization capacity of differential optical flow descriptors, as well as the capacity of MBH to differentiate between video content around Harris STIPs, as opposed to HOG and HOF descriptors. It should however be noted here that MBH performance comprises a complex multiple kernel combination of a horizontal MBHx and vertical MBHy component. In [16], a recognition accuracy of 73.7% is reported for a combination of Gist3D and Harris STIP + HOG/HOF descriptors. However, performance of the individual descriptors is

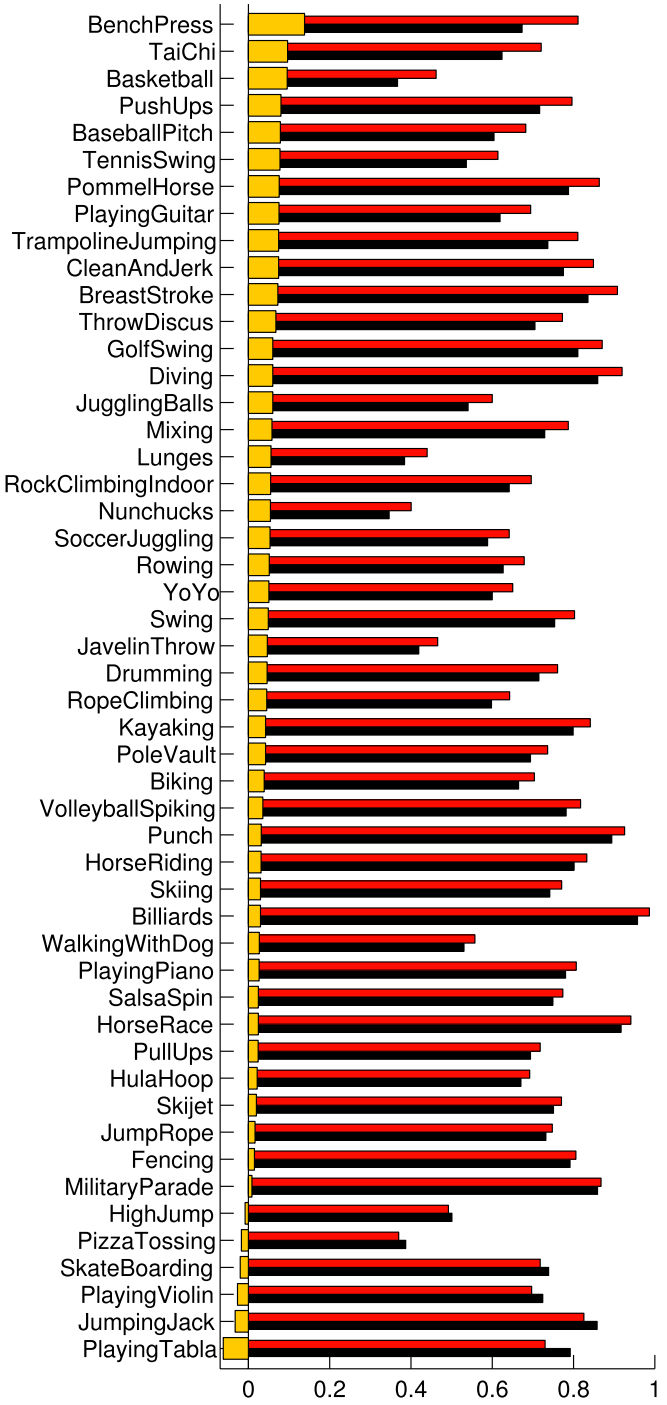

Fig. 9. Per-class recognition performances on UCF50 dataset. Color-STIP (IN-Gabor+$IC_{1,1}$) performance is depicted in red, intensity-STIP (I-Gabor+$I_{,1}$) in black and their difference in yellow (Color online).

at most 65.3%. In the recent work of Wang et. al. [25], trajectory-based HOG, HOF and MBH attain 68%, 68.2% and 82.2% respectively, while a multiple kernel combination yields state of the art performance of 84.5%. Finally, in [15] a result of 72.4% is obtained based on dense random sampling of HOG3D descriptors, whereas 83.3% is achieved with a multiple kernel combination of HOG, HOF, HOG3D and MBH descriptors.

We report a top performance of 72.9% for $IC_{1,1}$-HOG3D extracted around IN-Gabor STIPs. This result constitutes the best performing STIP-based approach to action recognition, while state of the art results are achieved by trajectory-based harvesting or dense sampling of MBH descriptors and multiple kernel modeling thereof.

1) Discussion on Per-Class Results: The results on UCF50 are further analyzed based on the per-category results in Fig. 9. The recognition performance for 44 out of 50 action categories is improved by using color. The largest improvement is observed for 'BenchPress'. The main reason for this is that the barbell weights are often (red) colored and thus render discriminative periodic motion patterns, see Fig. 10 for examples. Another influential factor is the associated typical indoor setting (gym), which often consists of solidly colored walls contrasting with the motion patterns in the foreground. Apart from that, we observe a large variety in terms of, for example, the specific background color or the clothing of the actors. Another action category with large recognition improvement is 'TaiChi'. We observe from corresponding examples that the activity is often performed outdoors on green grass by individuals wearing colorful clothes. Furthermore, it turns out that two 'TaiChi' video groups are composed of the same person performing the activity in the same pink clothes, which provides an obvious advantage to color based methods. A similar line of reasoning applies to the decreased recognition performance of 'PlayingTabla' activity, as one of the video groups contains grayscale samples only (in which all RGB channels are consequently identical). The subtraction of 'RGB' channels in the transformation to 'chromatic' opponent space in eq. 2 then yields $NULL$ channels. However, it is also possibly the case that the cast shadows of the fingers on the tabla exhibit discriminative motion patterns which may be better detected by an unnormalized (intensity-only) STIP detector. Another category for which intensity STIPs perform better is 'JumpingJack'. Also here, there is one video group containing essentially black/white footage which influences the results. We conclude from these observations that the usage of color for action recognition provides a performance boost in general, while the extremal result cases exhibit rather trivial characteristics.

I. Discussion on Entropy and Discriminative Power

Consider the descriptor with the highest entropy: $IC_{0,0}$-HOG3D. This is the best performing descriptor on UCF11, suggesting that high entropy is an indicator for discriminative power. On UCF50, however, $IC_{1,1}$-HOG3D is the best performing descriptor, which has considerably lower entropy compared to most other descriptors. When larger datasets exhibiting higher intra-class variability and lower inter-class variability are considered, it becomes more important for descriptors to be robust, as opposed to discriminative only. Another illustrative example of this phenomenon would be a raw pixel descriptor (list of pixel values) which typically has high entropy and is very discriminative but not at all robust. Another high-entropy descriptor is the 2-channel $IH_{0,0}$-HOG3D. This is remarkable at first sight because its dimensionality is lower than e.g. the 3-channel $IN_{0,0}$-HOG3D descriptor. That is, entropy is generally expected to increase along with dimensionality. Furthermore,

Fig. 10. Example frames from UCF50 dataset. The top row contains samples from the categories for which recognition performance based on color STIPs has improved the most over intensity STIPs. The bottom row shows examples from the 6 categories for which recognition performance has decreased. The samples are sorted from left to right based on the difference in recognition rates (Color online).

TABLE VI
SUMMARY OF BEST RECOGNITION RESULTS OVER ALL DATASETS

	UCF sports	UCF11	UCF50
# videos	150	1200	6700
# actions	10	11	50
Best Detector	IN	IN	IN
Best Descriptor	$IC_{0,0}$	$IC_{0,0}$	$IC_{1,1}$

the results on UCF sports show that IH descriptors perform worse than other descriptors in general which can be attributed to the instability of the hue representation for unsaturated colors resulting in high entropy in the extracted descriptor.

In conclusion, high descriptor entropy indicates either discriminative power or instability of the underlying representation. Discriminative power does not guarantee best performance because descriptor robustness becomes more important as the problem becomes more difficult.

VI. CONCLUSION

We have reformulated STIP detectors and descriptors to incorporate multiple photometric channels in addition to image intensities, resulting in color STIPs. The enhanced modeling of appearance results in an improved balance between photometric invariance and discriminative power, as chromaticity provides *more* information, based on which *better* representations are formed. Color STIPs are thoroughly evaluated and shown to significantly outperform their intensity-based counterparts for recognizing human actions on a number of challenging video benchmarks. In Table VI we show an overview of the best results over all datasets. The best detector is consistently IN, although differences between I and IN are small. Consistent across all results is the superior performance of descriptors extracted from the unnormalized opponent representation IC. Differences are observed between variations of the IC descriptor in terms of channel integration/concatenation and gradient orientation/direction, where the best descriptor choice depends on the difficulty and size of the dataset. For a small to moderate amount of visually relatively distinct categories such as in the UCF11 dataset, it is best to use a discriminative descriptor such as $IC_{0,0}$ (channel concatenation + gradient direction). For larger datasets such as UCF50 it is better to use the robust descriptor $IC_{1,1}$ (channel integration + gradient orientation), which has the additional advantage of low dimensionality.

REFERENCES

[1] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2010.
[2] G. J. Burghouts and J. M. Geusebroek, "Performance evaluation of local colour invariants," *Comput. Vis. Image Understand.*, vol. 113, no. 1, pp. 48–62, Jan. 2009.
[3] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ. Pittsburgh, PA, USA, 2009.
[4] P. Dollàr, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop VSPETS*, Oct. 2005, pp. 65–72.
[5] I. Everts, J. C. van Gemert, and T. Gevers, "Per-patch descriptor selection using surface and scene attributes," in *Proc. 12th ECCV*, 2012, pp. 172–186.
[6] I. Everts, J. C. van Gemert, and T. Gevers, "Evaluation of color STIPs for human action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 2850–2857.
[7] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. 19th BMVC*, 2008, pp. 275–285.
[8] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.
[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.
[10] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild'," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2009, pp. 1996–2003.
[11] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, 2012.
[12] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.
[13] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 1234–1241.
[14] S. Shafer, "Using color to separate reflection components," *Color Res. Appl.*, vol. 10, no. 4, pp. 210–218, 1985.
[15] F. Shi, E. Petriu, and R. Laganière, "Sampling strategies for real-time action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 2595–2602.
[16] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Mach. Vis. Appl.*, vol. 24, no. 7, pp. 1473–1485, Oct. 2012.
[17] F. Souza, E. Valle, G. Cámara-Chávez, and A. de Araújo, "An evaluation on color invariant based local spatiotemporal features for action recognition," presented at SIBGRAPI, Ouro Preto, Brazil, Aug. 2012.

[18] J. Stöttinger, A. Hanbury, N. Sebe, and T. Gevers, "Sparse color interest points for image retrieval and object categorization," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2681–2692, May 2012.
[19] J. Stöttinger, S. Zambanini, R. Khan, and A. Hanbury, "Feeval—A dataset for evaluation of spatio-temporal local features," in *Proc. 20th ICPR*, Aug. 2010, pp. 499–502.
[20] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, *et al.*, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3681–3688.
[21] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[22] J. van de Weijer, T. Gevers, and J. M. Geusebroek, "Edge and corner detection by photometric quasi-invariants," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 27, no. 4, pp. 625–630, Apr. 2005.
[23] J. van de Weijer, T. Gevers, and A. W. M. Smeulders, "Robust photometric invariant features from the colour tensor," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 118–127, Jan. 2006.
[24] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3169–3176.
[25] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, 2013.
[26] H. Wang, M. M. Ulla, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009, pp. 124.1–124.11.

Ivo Everts received the B.Sc. and M.Sc. degrees from the University of Amsterdam, The Netherlands, where he is currently pursuing the Ph.D. degree with the Intelligent Systems Lab Amsterdam, Faculty of Science. His research interests include computer vision, image processing, and pattern recognition.

Jan C. van Gemert received the B.Sc. degree from the Fontys University of Applied Sciences and the M.Sc. and Ph.D. degrees from the University of Amsterdam. He interned at MERL, Cambridge, MA, USA, and was a Visiting Researcher with the National Institute of Informatics, Tokyo, Japan. From 2008 to 2010, he was a Post-Doctoral Fellow with Cole Normale Suprieure, Paris, France, with Prof. J. Ponce. He founded Puzzual, a company in visual analytics. Currently, he is a Computer Vision Researcher with the University of Amsterdam with Prof. T. Gevers and Dr. C. Snoek. He teaches the computer vision master course and is supervising several Ph.D. students. His research interests include low-level visual features, image and video categorization, and action and object recognition. He has published over 30 papers where several are cited more than 100 times.

Theo Gevers (M'01) is a Full Professor of computer vision with the University of Amsterdam (UvA), Amsterdam, The Netherlands, and a Full Professor with the Computer Vision Center, Universitat Autnoma de Barcelona, Barcelona, Spain. He is a Founder and Chief Science Officer with Sightcorp, a spinoff of the Intelligent Systems Laboratory, UvA. His main research interests are in the fundamentals of image understanding, object recognition, and color in computer vision. He is interested in different aspects of human behavior, specifically in emotion recognition. He is the chair for various conferences and is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a program committee member for a number of conferences and an invited speaker at major conferences. He is a lecturer delivering post-doctoral courses given at various major conferences, including the IEEE Conference on Computer Vision and Pattern Recognition, the International Conference on Pattern Recognition, SPIE, and the Computer Graphics, Imaging, and Vision.