# Evaluation of Color STIPs for Human Action Recognition

Ivo Everts, Jan C. van Gemert and Theo Gevers
Intelligent Systems Lab Amsterdam
University of Amsterdam

## Abstract

*This paper is concerned with recognizing 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 highlights and shadows. Moreover, valuable information is neglected by discarding chromaticity from the photometric representation. These issues are addressed by Color STIPs. Color STIPs are multi-channel reformulations of existing intensity-based STIP detectors and descriptors, for which we consider a number of chromatic representations derived from the opponent color space. This enhanced modeling of appearance improves the quality of subsequent STIP detection and description. Color STIPs are shown to substantially outperform their intensity-based counterparts on the challenging UCF sports, UCF11 and UCF50 action recognition benchmarks. Moreover, the results show that color STIPs are currently the single best low-level feature choice for STIP-based approaches to human action recognition.*

## 1. 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 that work 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' [9]. 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 [12], and may build on low-level building blocks [18] which typically consider generic video representations based on local photometric features [6, 8, 23]. 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 and are consequently less scalable [12]. Low-level approaches are conceptually simple, easy to implement, sparse and efficient. Due to the local nature of features on which low-level approaches are based, they are naturally robust against geometric disturbances such as occlusion and viewpoint changes. Therefore, in this paper, we focus on low-level representations for recognizing human actions in video.

Low-level action recognition approaches are typically based on spatio-temporal interest points (STIPs) where image sequences are represented by descriptors extracted locally around STIP detections. These spatio-temporal feature detectors and descriptors typically use intensity-only representations of the video data and are therefore sensitive to disturbing illumination conditions such as shadows and highlights. More importantly, discriminative information is ignored by discarding chromaticity from the representation.

In a variety of image matching and object recognition tasks, color descriptors outperform intensity descriptors [2, 19] in the spatial (non-temporal) domain. We identify two benefits of adding color to the temporal domain. By using color, our approach can extract *more* temporal variations, since pure chromatic temporal transitions such as *e.g.*, red-green, or yellow-blue motion may not be visible in gray-scale. Further, because color is more discriminative, it allows for *better* estimation of motion and temporal variation. Where motion of colored objects may be ambiguous in gray-scale, color can be conclusive. Adding color to the temporal domain thus gives more information and it may improve the quality of the estimations.

In this paper, we propose to incorporate chromatic representations in the spatio-temporal domain. This comprises a reformulation of STIP detectors and descriptors for multi-channel video representations. For this, videos are represented in a variety of color spaces exhibiting different levels of photometric invariance. By this enhanced modeling of appearance, we aim to increase the quality (robustness and discriminative power) of STIP detectors and descriptors

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.367

for recognizing human activities in video. This is validated through a set of repeatability and recognition experiments on challenging video benchmarks. The results show that our color STIPs significantly outperform their intensity-based counterparts. Compared to existing work, color STIPs are favored over all other STIP-based approaches and perform competitively on the UCF50 dataset in comparison to the state of the art.

### 1.1. Related Work

In the spatial domain, multi-channel photometric invariant feature detectors [16, 20, 21] increase repeatability, entropy, and image categorization over intensity-based detections. For descriptors, multi-channel formulations [2, 19] propose various color SIFT variants where OpponentSIFT considerably improves 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. In contrast to other color-STIPS [15], we improve over standard baselines, use a well-founded representation model and we evaluate detectors and descriptors separately.

#### 1.1.1 Spatio-temporal Detectors

In the spatio-temporal domain, pioneering work by Laptev [7] extends the Harris function to 3D. Alternatively, there is the Gabor STIP detector of Dollàr *et al*. [4] which applies a Gabor filter along the temporal axis and is not based on differential image structure. The authors 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 [22]. Here, densely sampled points are tracked from frame to frame based on optical flow. As the method involves tracking and multi-scale optical flow computation, the associated computational complexity is typically higher than that of STIP-based approaches, but may compare favorably in terms of recognition rates. However, it is shown in [10] that motion-based descriptors are not scalable with respect to the number of action categories. This can be reasonably assumed to also hold for trajectory-based sampling of descriptors. In this paper, we focus on the sparser and more scalable STIP-based approach.

#### 1.1.2 Spatio-temporal Descriptors

Among the local spatio-temporal descriptors available in literature, the HOG3D descriptor [6] is well-suited for large scale video representation and multi-channel extensions. In contrast to *e.g*. HOG/HOF [8], MoSIFT [3] or MBH [22] descriptors, the HOG3D algorithm serves as an integrated and efficient approach, as it excludes optical flow which is computationally expensive [10]. 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 [23]. Therefore, we derive several multi-channel variants of the HOG3D descriptor and evaluate their performance for realistic human action recognition.

Another recently proposed video descriptor for human action recognition in web videos is Gist3D [14]. 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 only in combination with STIPs.

The works mentioned above comprise low/medium level approaches to action recognition. Higher level approaches such as Action Bank by Sadanand *et al*. [12] give excellent results on some datasets. However, such high-level approaches are typically not scalable. In contrast, low-level approaches are widely applicable, conceptually simple, sparse and exhibit reasonable computational complexity. Moreover, they may serve as powerful building blocks for higher level methods [18]. We contribute by considering a variety of photometric representations for STIP detection and description for enhancing low-level approaches to action recognition.

## 2. Photometric Representations

We model image formation by the dichromatic reflection model [13],

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

where a $RGB$ vector $\mathbf{f} = (R, G, B)^T$ 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, can be addressed by dividing-out the light source intensity $e$. Here, we consider the transformation of the $RGB$ image to the opponent color space [2, 5, 19, 20]

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

The transformation approximately decorrelates the image

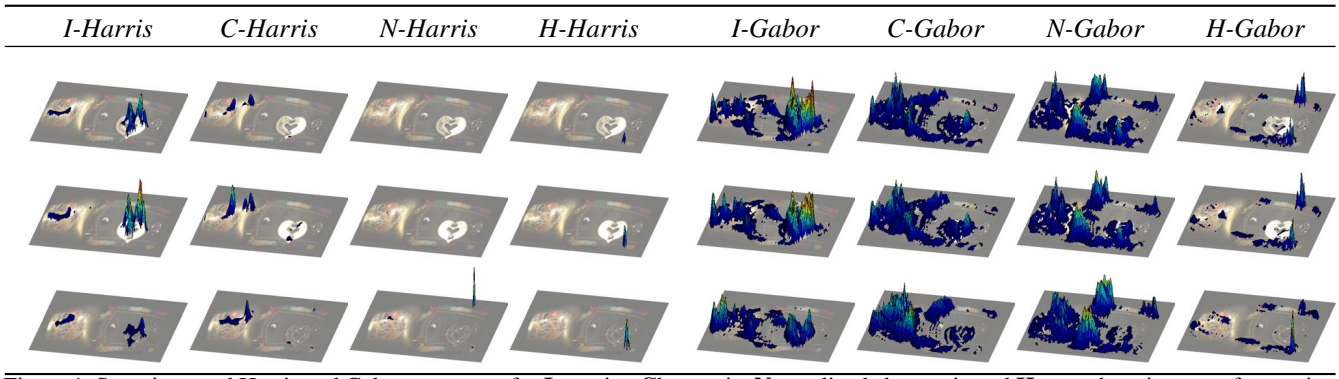


Figure 1. Superimposed Harris and Gabor responses for **I**ntensity, **C**hromatic, **N**ormalized chromatic and **H**ue on three images of a rotating object on which a strong highlight is present. The Harris energy function mainly responds to differential changes in the signal, whereas the Gabor function fires on general spatio-temporal fluctuations. Note the dampened response to the highlight in the invariant channels.

| | 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 | $I$ | $C$ | $N$ | $H$ |

Table 1. Photometric image representations. Chromatic combinations with the intensity channel yield *IC*, *IN* and *IH*.

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 canceled out in the formulations of $O_1$ and $O_2$, making the chromatic opponent components invariant to signal shifts such as those caused by (white) highlights.

**Shadow-shading**. The chromatic components are normalized by the 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 and 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 (1).

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 so as to weight them equally a-priori.

## 3. Multi-Channel STIP Detection

**Multi-channel Harris STIPs.** Harris STIPs are local maxima of the 3D Harris energy function based on the structure tensor [7]. A multi-channel formulation of the structure tensor has been developed in *e.g.* [21] which prevents opposing color gradient directions to cancel each other out. Here, we incorporate multiple channels in the spatio-temporal structure tensor [7].

The multi-channel volume $V$ consisting of $n_c$ channels is denoted by $V = (V^1, V^2, ..., V^{n_c})^T$. 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, $\sigma_o$ and $\tau_o$ are the spatial and temporal observation scales and $f^j : \mathbb{R}^3 \rightarrow \mathbb{R}$ is the imaging function of channel $j$.

Let $V_d = (V_d^1, V_d^2, ..., V_d^{n_c})^T, d \in \{x, y, t\}$ denote the per-channel partial Gaussian derivatives of the volume. The multi-channel spatio-temporal structure tensor is then defined by

$$S = g(\cdot; \sigma_i, \tau_i) * \begin{pmatrix} V_x \cdot V_x & V_x \cdot V_y & V_x \cdot V_t \\ V_y \cdot V_x & V_y \cdot V_y & V_y \cdot V_t \\ V_t \cdot V_x & V_t \cdot V_y & V_t \cdot V_t \end{pmatrix}, \quad (3)$$

where $\sigma_i$ and $\tau_i$ denote the spatial and temporal integration scale respectively. In figure (1), we illustrate the response per representation. Incorporating increasingly invariant photometric representations has a dramatic effect on the Harris energy. The highlight on the object surface triggers a strong response from the original $I$-based energy

functions. This effect is clearly dampened in the $C$ representation. However, the illumination reflected by the colored matte-shiny (left) object part still triggers response, as this reflection causes signal changes that are not captured 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 response except 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 definite by formulation. Hence, no additional care has to be taken to account for conflicting response signs between channels

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

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 figure (1), 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.

## 4. Multi-Channel STIP Description

The HOG3D [6] descriptor 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 $\boldsymbol{g}$,

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

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 over every grid cell and concatenated to form the final descriptor [6].

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

Table 2. 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 ($1D$, 360 in this paper), based on which we derive the dimensionality of the other descriptor variants.

Chromaticity is incorporated in the HOG3D descriptor by considering the representations from section (2) in a multi-channel formulation of the gradient vector $\boldsymbol{g}$ in eq. (5). We evaluate the standard practice of concatenation of the per-channel descriptors [2, 5, 19]:

$$\boldsymbol{g'} = \{\boldsymbol{g}^j\}, j = 1, ..., n_c. \tag{6}$$

We also evaluate a single gradient variant where we prevent the effect of opposing color gradient directions by using tensor mathematics. In tensors, opposing directions reinforce each other by summing the gradient *orientations* as opposed to their *directions* [21],

$$\boldsymbol{g''} = \sum_{j=1}^{n_c} \boldsymbol{g}^j \cdot \boldsymbol{g}^j. \tag{7}$$

This formulation of the gradient defines half of the full sphere of directions which is one of the HOG3D flavors in [6]. 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*. This 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 (2).

## 5. Experiments

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

### 5.1. 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.* [6].

**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 $\tau_o$ is in conflict with *e.g.* [23], 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. We have found this to be satisfactory in terms of discriminative power and computational load. Specifically, for computing 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.* [23, 8], we do not perform STIP scale selection because of its associated high computational costs and decreased recognition performance [7].

**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 [6]. 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.* [23]) 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 3x3x2 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 (2). 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 $I$ is 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 [6], because this affects the very 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 usually quite influential and 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 [6] 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 [23]. 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. A SVM is trained based on the $\chi^2$ 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 the amount of videos and the number of STIPs).

## 5.2. Datasets

We measure STIP repeatability on videos from the **FeEval** dataset [17]. This dataset consists of 30 videos taken from television series, movies and lab recordings. Every 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 television series up to the first occurring shot boundary. That is, we do not aim 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.

For an in-depth evaluation of detector and descriptor settings, we use the **UCF sports** dataset [11]. 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 authors of [23] have kindly provided us with an exact copy of the dataset as used in their experiments. The best performing experimental settings are applied to **UCF11** [9] which has 11 human actions in 1200 videos, and its superset **UCF50** [10] with 50 human action classes in about 6700 videos. 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].

| | $I$ | $IC$ | $IN$ | $IH$ | $C$ | $N$ | $H$ |
|---|---|---|---|---|---|---|---|
| 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% |

Table 3. STIP repeatability for multi-channel Harris and Gabor detectors based on the considered photometric representations.

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

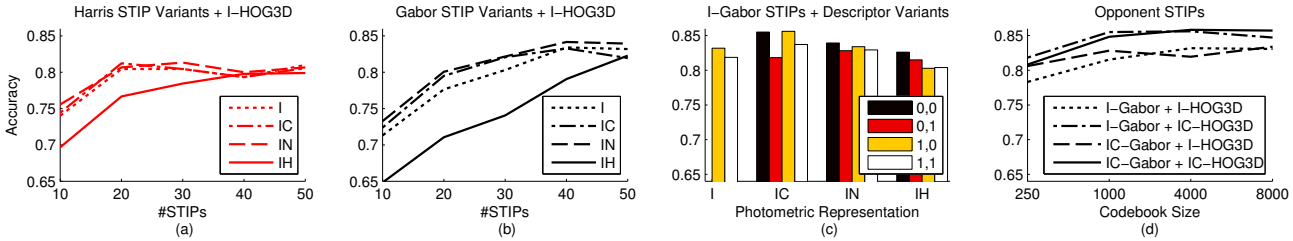


Figure 2. Recognition performance on the UCF sports dataset per photometric representation for varying amounts of Harris (a) and Gabor (b) STIPs. Influence of the photometric representations on descriptor variants (c). Combinations of the top-performing $IC$-Gabor STIPs and $IC_{1,0}$-descriptors for varying codebook sizes (d).

### 5.3. Color STIP Detector Repeatability

We poll the detectors for an average amount 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 (3).

Harris STIPs are much more stable than Gabor STIPs. Nonlinear differential spatio-temporal signal changes are more distinctive than temporal fluctuations only. The behavior of the detector in different photometric representations are in line with figure (1). 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 (save marginal improvements for the $IC$ representation). Moreover, the $IH$ representation attains much lower repeatability scores than $I$. The reason for this is that, as disturbing conditions such as highlights and shadows are effectively ignored, so is spatio-temporal image structure on which stable STIPs are detected. Adding $C$ or $N$ to the intensity $I$ basically leaves the repeatabililty unaltered on this dataset. However, the STIP discriminability experiments will show that adopting these representations does result in different recognition scores.

From here on, the pure chromatic representations are discarded from the experimental batch due to the associated lack of discriminative power and we focus only on $I$, $IC$, $IN$, $IH$.

### 5.4. 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 detector. Direction-based intensity HOG3D ($I_{,0}$) descriptors are extracted around multi-channel STIP detections (*i.e.* the descriptor representation is fixed). 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 figures (2a,b).

We first validate our implementation by comparing recognition accuracies with the evaluations reported on intensity in [23]. 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, [23] reports an accuracy of 82.9% for 44 STIPs. This is comparable to our performance of 83.4% for 40 STIPs.

From figures (2a,b) it stands out 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. Harris STIPs appear more discriminative than Gabor STIPs for relatively small amounts of detections. This relative performance difference reverses as more STIPs are considered. The reasons for this are related to sparsity, distinctiveness and scale.

Considering Harris STIPs in figure (2a), using 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 by $IN$. The primary characterization of Harris STIPs in terms of distinctiveness and sparseness 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.

For the multi-channel Gabor detector in figure (2b) higher quality 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)).

### 5.5. Color STIP Descriptor Discriminability

For the following action recognition experiments on the UCF sports dataset, descriptors are extracted around Gabor STIPs as these have shown superior recognition per-

formance over Harris STIPs in figure (2a,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 figure (2c).

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 2). 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 to multiple channels and increased dimensionality. This is also the most widely spread approach to multi-channel descriptors, *e.g.* [2, 19, 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 [23].

We conduct a final investigation on the codebook size. We consider 'OpponentSTIP' 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 figure (2d). 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 [12] 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 [23], 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 88.2% is achieved in [22]. Compared to this, our STIP-based approach does a good job considering that it outperforms all reported individual features on UCF sports.

### 5.6. UCF11 & UCF50

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 (4).

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. This compares to the trajectory-based harvesting of HOG and HOF features in [22], for which 74.5% and 72.8% is achieved respectively. However, they report a superior performance of 83.9% for MBH. In our case, adding chromaticity increases the recognition accuracies substantially where best performance is achieved by the direction-based $IC$ descriptors: 78.4% for $IC_{1,0}$ on $IC$-Gabor STIPs and 78.6% 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.

Interestingly, best performance on UCF50 is achieved by orientation-based descriptors. As the number of categories increases, descriptor robustness becomes more important. We observe a baseline result of 68.8% for $I_{,1}$. This is substantially higher than the results reported in [12] for Action Bank (57.9%) and Harris STIP + HOG/HOF (47.9%) (see table (5) for an overview of recent results on UCF50). We conclude from this that the Action Bank method is not scalable, and probably suffers from increased amounts of geometric variations. As for Harris STIP + HOG/HOF, we conclude that the high degree of distinctiveness of spatio-temporal corners limits generalization capacity. A performance of 76.9% is reported in [10] for a combination of scene context and spatio-temporal descriptors. Here, the best performing single spatio-temporal descriptor is MBH [22], which achieves 71.9%. This shows the generalization capacity of differential optical flow descriptors. In [14], a recognition accuracy of 73.7% is reported for a combination of Gist3D and STIP (HOG/HOF) descriptors. However, their performance of the individual descriptors are at most 65.3%.

We report a top performance of 72.9% for $IC_{1,1}$-

| | | $I_{.,0}$ | $IC_{1,0}$ | $IC_{0,0}$ | $IN_{1,0}$ | $IN_{0,0}$ | $I_{.,1}$ | $IC_{1,1}$ | $IC_{0,1}$ | $IN_{1,1}$ | $IN_{0,1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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% |

Table 4. 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.

| Ref. | Description | % |
|---|---|---|
| [10] | Scene context + STIP(MBH) | 76.9% |
| | Scene Context | 47.6% |
| | STIP(MBH) | 71.9% |
| [14] | Gist3D + STIP(HOG/HOF) | 73.7% |
| | Gist3D | 65.3% |
| | STIP(HOG/HOF) | 54.3% |
| [12] | Action Bank | 57.9% |
| | STIP(HOG/HOF) | 47.9% |
| Here | Color STIP(HOG3D) | 72.9% |

Table 5. Recent UCF50 results available in literature.

HOG3D extracted around $IN$-Gabor STIPs. This result is highly competitive compared to the state of the art, considering that it involves only a single descriptor type.

## 6. Conclusion

We have reformulated existing STIP detectors and descriptors to incorporate multiple photometric channels, resulting in Color STIPs. This enhanced modeling of appearance results in higher quality detections and descriptions. 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 general, best results are obtained based on unnormalized opponent color representations.

## References


[1] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *PAMI*, 2010. 5

[2] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 2009. 1, 2, 4, 7

[3] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos, 2009. 2

[4] P. Dollàr, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV-VSPETS*, 2005. 2, 4

[5] I. Everts, J. C. van Gemert, and T. Gevers. Per-patch descriptor selection using surface and scene attributes. In *ECCV*, 2012. 2, 4, 7

[6] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 1, 2, 4, 5

[7] I. Laptev. On space-time interest points. *IJCV*, 2005. 2, 3, 5

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2, 5

[9] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009. 1, 5

[10] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *MVAP*, 2012. 2, 5, 7, 8

[11] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 5

[12] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012. 1, 2, 7, 8

[13] S. Shafer. Using color to separate reflection components. *Color research and applications 10*, 1985. 2

[14] B. Solmaz, S. M. Assari, and M. Shah. Classifying web videos using a global video descriptor. *MVAP*, 2012. 2, 7, 8

[15] F. Souza, E. Valle, G. Cámara-Chávez, and A. d. A. Araújo. An evaluation on color invariant based local spatiotemporal features for action recognition. In *IEEE SIBGRAPI*, 2012. 2

[16] J. Stöttinger, A. Hanbury, N. Sebe, and T. Gevers. Sparse color interest points for image retrieval and object categorization. *TIP*, 2012. 2

[17] J. Stöttinger, S. Zambanini, R. Khan, and A. Hanbury. Feeval - a dataset for evaluation of spatio-temporal local features. In *ICPR*, 2010. 5

[18] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012. 1, 2

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010. 1, 2, 4, 7

[20] J. van de Weijer, T. Gevers, and J. M. Geusebroek. Edge and corner detection by photometric quasi-invariants. *PAMI*, 2005. 2

[21] J. van de Weijer, T. Gevers, and A. W. M. Smeulders. Robust photometric invariant features from the colour tensor. *TIP*, 2006. 2, 3, 4

[22] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2, 7

[23] H. Wang, M. M. Ulla, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 1, 2, 5, 6, 7