Color-Aware Local Spatiotemporal Features
for Action Recognition

Fillipe Souza[1], Eduardo Valle[2], Guillermo Chávez[3], and Arnaldo de A. Araújo[1]

[1] NPDI Lab, DCC/UFMG, Belo Horizonte/MG, Brazil
[2] RECOD Lab, IC/Unicamp, Campinas/SP, Brazil
[3] ICEB/UFOP, Ouro Preto/MG, Brazil
`fdms18@gmail.com, mail@eduardovalle.com,`
`gcamarac@gmail.com, arnaldo@dcc.ufmg.br`

Abstract. Despite the recent developments in spatiotemporal local features for action recognition in video sequences, local color information has so far been ignored. However, color has been proved an important element to the success of automated recognition of objects and scenes. In this paper we extend the space-time interest point descriptor STIP to take into account the color information on the features' neighborhood. We compare the performance of our color-aware version of STIP (which we have called HueSTIP) with the original one.

Keywords: Color invariance, spatiotemporal local features, human action recognition.

1 Introduction

In this work we provide a discussion on the role of spatiotemporal color features for human action recognition in realistic settings. Color is a prominent feature of the real world scenes and objects. Not surprisingly, it has become a powerful tool in automated object recognition [1] [2] [3]. However, color has not yet been given its deserved importance in the universe of unconstrained action recognition.

Several spatiotemporal local feature descriptors and detectors have been proposed and evaluated in action recognition. Detectors rely commonly on a measure function (or response function) to locate interest regions. Those local regions (also called patches) can be described in terms, for example, of histograms of gradient orientations and optical flow. Laptev [4] presented a spatiotemporal extension of the Harris-Laplace corner detector proposed by Mikolajczyk and Schmid [5]. Spatiotemporal corners are found when strong intensity variations over the spatial and temporal domains occur simultaneously. This method has proved efficient for action recognition in controlled datasets such as the KTH dataset [6]. Dollár et al. [7] proposed a spatiotemporal detector based on temporal Gabor filters that considers only local variations having periodic frequency components. Another spatiotemporal interest point detector was designed in [8] by Willems et al. This detector uses the Hessian determinant as a saliency measure and 3D convolution approximations by box-filters in order to find regions

of interest. Here we will only provide a formal discussion of the spatiotemporal local feature detector used in this work, the one proposed in [4].

To improve the discriminative power and illumination invariance of local features to object recognition and image categorization, a set of color descriptors for spatial local features was proposed in [3] by van de Sande et al. for static images. The distictiveness of their color descriptors was evaluated experimentally and their invariant properties under illuminations changes were analyzed. They derived different color descriptors, including combinations with the intensity-based shape descriptor SIFT [9]. Van de Weijer et al. [2] had already proposed color histograms providing robustness to photometric and geometrical changes, photometric stability and generality. Their work was the basis for some of the descriptors developed in [3], including our own.

The most important contribution of this work is the combination of the work by van de Weijer et al. [2] and Laptev [4] to propose the use of color descriptors to describe local spatiotemporal interest points (features). To this end, we have produced preliminary results on the matter, such that only information from the hue channel (of the HSI color system) is evaluated as the source of color data for the description of space-time interest points. The second contribution is an analysis of the proposal on the prevailing application of automated recognition of human actions (on a challenging dataset, the Hollywood2 [10]). We make a performance comparison between which has been considered as standard in the literature and our proposal. We then discuss the results clarifying the cases of success and failure brought by the addition of color information.

The rest of this paper is organized as follows. In section 2 we are concerned with the formal description of the spatiotemporal interest point detector used. Further, the details on the color descriptors are presented in section 3. The experiments and their results are discussed in section 4 and section 5 concludes the work.

2   Spatiotemporal Interest Points

Laptev [4] proposed a differential operator that checks for local extremas over spatial and temporal scales at the same time. Extrema in specific space-time locations refer to particular patterns of events. This method is built on the Harris [11] and Förstner [12] interest point operators, but extended to the temporal space. Essentially, as a corner moves across an image sequence, at the change of its direction an interest point is identified. Other typical situations are when image structures are either split or unified. For being one of the major elements in this work, a few details and mathematical considerations on the detector design are presented next.

Many interest events in videos are characterized by motion variations of image structures over time. In order to retain those important information, the concept of spatial interest points is extended to the spatio-temporal domain. This way, the local regions around the interest points are described with respect to derivatives in both directions (space and time).

At first, the selection of interest point in the spatial domain is described. The linear scale-space representation of an image can be mathematically defined as $L^{sp} : R^2 \times R_+ \mapsto R$, which is the convolution of f^{sp} with g^{sp}, where $f^{sp} : R^2 \mapsto R$ represents a simple model of an image and g^{sp} is the Gaussian kernel of variance σ_l^2. Then,

$$L^{sp}(x, y; \sigma_l^2) = g^{sp}(x, y; \sigma_l^2) * f^{sp}(x, y), \tag{1}$$

and

$$g^{sp}(x, y; \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp(-(x^2 + y^2)/2\sigma_l^2). \tag{2}$$

Localizing interest points means to find strong variations of image intensities along the two directions of the image. To determine those local regions, the second moment matrix is integrated over a Gaussian window having variance σ_i^2, for different scales of observation σ_l^2, which is written as the equation:

$$\mu^{sp}(.; \sigma_l^2, \sigma_i^2) = g^{sp}(.; \sigma_i^2) * ((\nabla L(.; \sigma_l^2))(\nabla L(.; \sigma_l^2))^T)$$
$$= g^{sp}(.; \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp}L_y^{sp} \\ L_x^{sp}L_y^{sp} & (L_y^{sp})^2. \end{pmatrix} \tag{3}$$

The descriptors of variations along the dimensions of f^{sp} are the eigenvalues of Equation 3: λ_1 and λ_2, with $\lambda_1 \leq \lambda_2$. Higher values of those eigenvalues are a sign of an interest point and generally lead to positive local maxima of the Harris corner function, provided that the ratio $\alpha = \lambda_2/\lambda_1$ is high and satisfies the constraint $k \leq \alpha/(1 + \alpha)^2$:

$$H^{sp} = \det(\mu^{sp}) - k.trace^2(\mu^{sp})$$
$$= \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2. \tag{4}$$

Analogously, the procedure to detect interest points in the scape-time domain is derived by rewriting the equations to consider the temporal dimension. Thus, having an image sequence modeled as $f : R^2 \times R \mapsto R$, its linear representation becomes $L : R^2 \times R \times R_+^2 \mapsto R$, but over two independent variances σ_l^2 (spatial) and τ_l^2 (temporal) using an anisotropic Gaussian kernel $g(.; \sigma_l^2, \tau_l^2)$. Therefore, the complete set of equations for detecting interest points described in [4] is the following.

$$L(.; \sigma_l^2, \tau_l^2) = g(.; \sigma_l^2, \tau_l^2) * f(.), \tag{5}$$

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}}$$
$$\times \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/\tau_l^2), \tag{6}$$

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2. \end{pmatrix} \tag{7}$$

$$H = \det(\mu) - k.trace^3(\mu)$$
$$= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3, \tag{8}$$

restricted to $H \geq 0$, with $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$, and subject to $k \leq \alpha\beta/(1 + \alpha + \beta)^3$

3   Local Features

Given a local interest region denoted by a spatiotemporal interest point, 3D local features accounting for appearance (histograms of oriented gradient) and motion (histograms of optical flow) are computed by using information from the neighborhood at (x, y, t). A spatiotemporal volume is sliced into $n_x \times n_y \times n_t$ 3D cells, in particular, $n_x = n_y = 3$ and $n_t = 2$. For each cell 4-bin histograms of gradient orientations (**HoG**) and 5-bin histograms of optical flow (**HoF**) are calculated, normalized and concatenated (**HoGHoF**, used by STIP [4]).

3.1   Color Descriptor

In this section, the hue histogram based color descriptor is roughly described. From the work in [2], the hue calculation has the form:

$$hue = \arctan\left(\frac{\sqrt{3}(R - G)}{R + G - 2B}\right). \tag{9}$$

It is known that, in the HSI color space, the hue value becomes unstable as it approaches the grey axis. In attempt to atenuate this problem, van de Weijer et al. [2] analyzed the error propagation in the hue transformation and verified the inverse proportionality of the hue certainty to the saturation. This way, the authors demonstrated that the hue color model achieves robustness by weighing the hue sample by the corresponding saturation, which is given by Equation 10:

$$sat = \sqrt{\frac{2(R^2 + G^2 + B^2 - RG - RB - GB)}{3}}. \tag{10}$$

To construct the hue histogram, we calculate the bin number to which the hue value (of the spatiotemporal volume) belongs with $bin = hue * 36/2\pi$. Then, at the position bin of the histogram the saturation value is accumulated. Before incrementing the histogram bin with a given amount of saturation, the saturation is weighed by a corresponding value of a Gaussian mask having the size of the spatiotemporal volume. The size and values forming the spatiotemporal Gaussian mask will vary according to the spatial and temporal scales of the interest point. The computed hue histogram will be further concatenated to the HoGHoF feature vector and this combination will be called **HueSTIP**.

4   Experiments

In our experiments, we investigated the power of the spatiotemporal local features containing color information for action recognition. This section describes the experimental setup followed by the analysis of the obtained results.

4.1   Dataset

We wanted to evaluate the performance of the descriptors for human action recognition in natural scenarios. Therefore, the Hollywood2 dataset [10] was a natural choice. This dataset is composed of 12 action classes: answering phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, standing up (see Figure 1). Videos were collected from a set of 69 different Hollywood movies, where 33 were used to generate the training set and 36 the test set. Action video clips were divided in three separate subsets, namely an automatic (noisy) training set, a (clean) training set and the test set. We only used the clean training set containing 823 samples and the test set containing 884 samples.

Fig. 1. Illustration of the Hollywood2 dataset containing human action from Hollywood movies

4.2   Bag-of-Features Video Representation

When spatiotemporal local features are extracted, they only provide a very local and disconnected representation of the video clips. One way to give a more meaningful representation is to use the *bag-of-features* (**BoF**) approach, which has been successfully applied to many applications of video analysis [3] [13]. Using BoF requires the construction of a vocabulary of features (or visual vocabulary). Although this is commonly accomplished by using k-means, it is well known that for very high-dimensional spaces, simple clustering algorithms perform badly, and thus a reasonable and efficient choice is just to select a random sample to form the visual vocabulary: this saves computational time and achieves comparable results [14]. The vocabulary size was set to 4000 since this number has empirically demonstrated good results and is consistent with the literature [13] [10]. At this point, spatiotemporal local features of a video clip are assigned to the closest visual word of the vocabulary (we use Euclidean distance function), producing a histogram of visual words. This histogram of visual word frequencies now accounts for the new video representation.

4.3   Classification

To classify the videos, we have used Support Vector Machines (SVM), using the LibSVM [15] implementation. Since our aim is to highlight the performance of the descriptors, we have chosen to simplify the classifier by using linear kernels

(experiments with more complex kernels were performed with comparable results). SVM being a binary classifier, LibSVM implements multi-classification by the one-to-one method, which creates $n(n-1)/2$ binary classifiers (where n is the number of classes) and applies a majority voting scheme to assign the class of an unknown element.

4.4   Experimental Protocol

1. Extract local features of the whole dataset (using both descriptors, HueSTIP and STIP),
2. Build the visual vocabularies, one for each feature type (HueSTIP or STIP),
3. Assemble the histograms of visual words representing each video clips of the dataset,
4. Learn the classifiers (one for each feature type) of the clean training set using SVM, in which the training and test samples are already separately available, as described in 4.1,
5. Classify the samples of the test dataset.

4.5   Results and Discussion

Table 1 evaluates the performance of both descriptors, STIP and HueSTIP, for the human action recognition task. It shows that there exists a gain in using color information for the classification of specific actions. Especially, half of the classes had the best performance achieved by HueSTIP, namely *AnswerPhone*, *FightPerson*, *HugPerson*, *Run*, *SitDown*, and *StandUp*. This increased performance brought by HueSTIP may have come either from information retrieved from parts of the objects of interest in the foreground or the background scenarios that is usually ignored by traditional shape or motion descriptors but gains meaning as the color description is considered.

Performance improvements achieved by HueSTIP for the *AnswerPhone* class can be justified by the color information from the background describing the indoor scenario in which this action usually takes place (see Figure 2 A) for an illustration of this fact), which are very similar among videos of this type of action. This is also the case for the *SitDown* class, as can be seen in Figure 2 B). For the *FightPerson* class, we have that in situations involving aggressive behaviors, the presence of blood can be expected, which can be an important aspect of the scene if color information is taken into account. Also, it is clear by the images in Figure 2 C) that features extracted from the scenes of violence acts where the camera mainly focuses on the abrupt moving of faces and arms of the actors will hold information of skin color in their descriptions, which together with features of optical flow and gradients may help to narrow the meaning of fight video's feature vectors. Regarding the class *Run* and *StandUp*, color information from the outdoor scenario might be useful.

However, for many other classes, the addition of color information actually results in losses. This is somewhat intuitive in *DriveCar* and *GetOutCar*, where the color variablity of cars acts more as a confusion than a help. The huge loss in
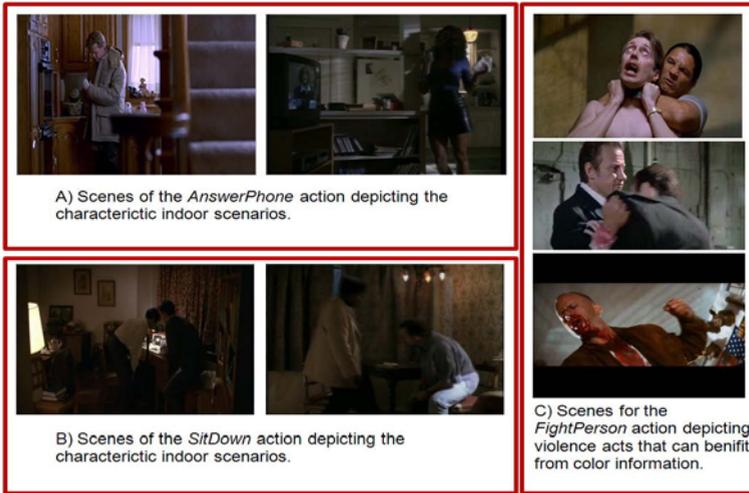
Fig. 2. Samples of scenes depicting situations that explain some of the results

Table 1. This table reports the accuracy rates given by each feature algorithm at the recognition of the twelve human actions depicted by videos of the Hollywood2 dataset

Action	HueSTIP	STIP	Action	HueSTIP	STIP
AnswerPhone	**12.5%**	9.4%	*HugPerson*	**18.2%**	12.1%
DriveCar	71.6%	**76.5%**	*Kiss*	38.8%	**49.5%**
Eat	45.5%	**57.6%**	*Run*	**58.9%**	57.5%
FightPerson	**68.6%**	62.9%	*SitDown*	**45.4%**	41.7%
GetOutCar	8.8%	**19.3%**	*SitUp*	0.0%	0.0%
HandShake	6.7%	**8.9%**	*StandUp*	**54.1%**	51.4%

performance in classes like *Eat* and *Kiss*, however was somewhat unexpected and reveal the weakness of using the same neighborhood for extracting the optical flow and color information. It is important to note that the lower performance in those cases (using HueSTIP) might have been caused by the difference in the density of points if compared with STIP, which had a greater number of points extracted due to a different parametrization.

5 Conclusion

We consider that HueSTIP has shown promising results for a preliminary work: the experiments show it can improve classification rates of actions, but that this improvement tends to be very class-dependent.

We are currently working on some of its interesting issues, especially the reasons why its performance is so unexpectedly low in a few classes. We suspect that using the same feature detector for STIP and HueSTIP might give the

former an unfair advantage for the classes where the interesting color phenomena happens at different scales than interesting grayscale phenomena.

References

1. Gevers, T., Stokman, H.: Robust histogram construction from color invariants for object recognition. PAMI 26, 113–117 (2004)
2. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 334–348. Springer, Heidelberg (2006)
3. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI 32(9), 1582–1596 (2010)
4. Laptev, I.: On space-time interest points. IJCV 64(2-3), 107–123 (2005)
5. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV 60(1), 63–86 (2004)
6. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC, p. 127 (September 2009)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (October 2005)
8. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
10. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
11. Harris, C., Stephens, M.: A combined corner and edge detector. In: 4th Alvey Vision Conf., pp. 147–152 (1988)
12. Forstner, W., Gulch, E.: A fast operator for detection and precise location of distinct points, corners and centres of circular features. In: ISPRS, pp. 281–305 (1987)
13. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (June 2008)
14. Viitaniemi, V., Laaksonen, J.: Experiments on selection of codebooks for local image feature histograms. In: Sebillo, M., Vitiello, G., Schaefer, G. (eds.) VISUAL 2008. LNCS, vol. 5188, pp. 126–137. Springer, Heidelberg (2008)
15. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software http://www.csie.ntu.edu.tw/~cjlin/libsvm