An Evaluation on Color Invariant Based Local Spatiotemporal Features for Action Recognition

Fillipe Souza*, Eduardo Valle†, G. Cámara-Chávez‡ and Arnaldo de A. Araújo*
*Department of Computer Science
Federal University of Minas Gerais, Belo Horizonte - MG, Brazil
Email: {fdms,arnaldo}@dcc.ufmg.br
†University of Campinas, Campinas - SP, Brazil
Email: dovalle@dca.fee.unicamp.br
‡Federal University of Ouro Preto, Ouro Preto - MG, Brazil
Email: guillermo@iceb.ufop.br

***Abstract*—Despite recent advances in the design of features to improve automated human action recognition, color information has so far been overlooked. Nevertheless, color has been proven an important element to the success of automated recognition of objects/scenes and segmentation. For object and scene recognition in static images, robustness to photometric variations has been achieved by describing local regions of spatial interest points in terms of color invariance properties. Such robustness was built on the dichromatic refection model proposed by Shafer. Thus, we extended the space-time interest point detector to incorporate color invariance properties in the feature extraction procedure. We were certain that color could contribute to the distictiveness of some classes. Additionally, in some cases, objects of interest are exhibited in a way that color appears as an essential element in describing the event. We evaluate the performance of the family of color-based STIPs in different application contexts: human actions from movies, violence and pornography. In the former, accuracy rates were improved in more than half of the action classes. In the pornography case, we found that the proposals are the best to increase reliability in critical applications such as digital forensics.**[1]

***Keywords*-color invariants; spatiotemporal features; violence detection; pornography detection; human actions**

I. Introduction

Multimedia processing has now evolved enough to allow high-level semantic filters in practical applications. Demands of this nature come from many social contexts, including security, health, and quality of communication, to name a few. In the scope of security, much research has focused on i) automating the analysis of surveillance videos to help human operators detect suspicious and abnormal behaviors where safety and quietness must be prevailed [1], and ii) developing biometric systems for automated authentication of authorized people. Examples regarding the quality of communication include work on i) automated interpretation of sign language sentences to facilitate learning and communication of hearing/speech-impaired people [2], and ii) filtering of inappropriate content (e.g., pornography and violence scenes) for a specific audience [3], [4], [5], [6], [7], [8], [9], [10], [11], [12].

When videos are the target media of analysis, it is a common practice to hypothesize that the contained motion information retains important latent patterns that could distinguishly describe the content of the video. Recently, local spatiotemporal features [13] have been proposed to describe motion patterns of objects performed in videos. Those features have been successfully applied in the context of human action recognition, which is the principal focus of several applications. Once motion patterns are localized, they are given a low level representation by histograms of oriented gradients and histograms of optical flow. The former is expected to provide shape description, whereas the latter is responsible for the inference of the apparent motion.

Intuitively, because objects are moving across the scenes, we consider trajectories and motion dynamics of parts of the objects as important elements describing the events in a scene. On the other hand, it is not evident that color information would be as instructive as motion patterns for distinctiveness. There is, however, much work supporting the use of color to improve recognition of objects and segmentation in still images [14], [15], where many robust color descriptors were proposed. In contrast, applications that rely on video processing have overlooked the role of color information to describe object actions and events in videos.

Motivation: Boosting the performance rates of human action recognition is an ultimate goal of long pursuit. If successfully accomplished, a diverse set of relevant applications would have arguably half of its problems overcome. There is a still a lot of aspects that remain to be explored and one of the most exiciting ones is to design the appropriate set of features. Most of the time this is dependent on the application and a lot of discussion revolves around whether or not only a subset of several types would not be sufficient for several of the tasks. Here we investigate the role of color information as low level features to represent different human actions, which had not received any attention by the time this work was done, to the best of our knowledge.

[1]***This work was a master's thesis developed at the UFMG as a partial fulfillment to confer the degree of MSc. in Computer Science to Fillipe D. M. de Souza. This work was completely developed by the student Fillipe de Souza while advised by Prof. Arnaldo de A. Araújo, and co-advised by Prof. Eduardo Valle and Prof. Guillermo C.-Chávez***

Contributions: This work's contribution is three-fold: First, we extended the space-time interest point detector (STIP) to incorporate color information (using the *normalized*-RGB color system, which will be later referred to as *rgb*) at the detection phase, which we have called the *ColorSTIP*. Secondly, we considered the combination of color histograms (based on the saturation-weighed hue channel) with the original STIP's descriptors to describe support regions of the interest points, which we named as *HueSTIP*. Finally, we conducted a performance analysis of the proposals as feature solutions to represent human actions in unconstrained scenarios for different application contexts: filtering of unwanted/prohibited content and content-based video indexing and retrieval.

II. The Proposals

The STIP's procedure to detect interest points in the scape-time domain is an extension of the Harris corner detector's equations to take into account the temporal dimension. Given an image sequence modeled as $f : R^2 \times R \mapsto R$, a linear representation of this sequence can be described by $L : R^2 \times R \times R^2_+ \mapsto R$. The detection of space-time interest points is described in [16] is the following set of equations:

$$L(.;\sigma_l^2) = g(.;\sigma_l^2,\tau_l^2) * f(.), \tag{1}$$

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

$$\mu = g(.;\sigma_i^2,\tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix}, \tag{3}$$

and

$$\begin{aligned} H &= \det(\mu) - k.trace^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3, \end{aligned} \tag{4}$$

where $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$, and $k \leq \alpha\beta/(1+\alpha+\beta)^3$. σ_l^2 (spatial) and τ_l^2 (temporal) are two independent variances used to describe the anisotropic Gaussian kernel $g(.;\sigma_l^2,\tau_l^2)$.

Our work extends the Laptev's space-time interest point detector to incorporate color information in ultimately two ways:

1) First, instead of the intensity gray-scale channel, the three *rgb* color channels were input to the interest point operator. (ColorSTIP)
2) Second, the support local space-time regions were also described in terms of histograms of *hue*; later combining it with the default HoG-HoF feature histogram used by STIP. (HueSTIP)

An intuitive, yet interesting, adaptation was to combine both proposals which we called Hue-ColorSTIP. As follows, we give a discussion on the two schemes used to yield the family of color-based STIPs.

A. Proposal 1: ColorSTIP

In this version, we try to detect space-time interest points also robust to photometric variations. To this end, for each frame, we replace the gray-scale channel by the three channels from the *rgb* system. So, the Gaussian derivatives are applied to each color channel at all directions ($x, y,$ and t). In the next step, the final color-based Gaussian derivative at each dimension is obtained by the summation of its Gaussian derivatives from each color channel, which becomes

$$L_x^{rgb} = L_x^r + L_x^g + L_x^b,$$

$$L_y^{rgb} = L_y^r + L_y^g + L_y^b,$$

$$L_t^{rgb} = L_t^r + L_t^g + L_t^b.$$

The *rbg*-based second moment matrix μ^{rgb} from the space-time extension of the Harris-Laplace corner detector will then be given by

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

while the rest of the method continues the same. By doing this, the original STIP starts to hold properties that it lacked before. The properties to which we refer stem from the photometric invariant properties held by the *rbg* system, namely lighting geometry, viewpoints and illumination intensity [15]. With those additional characteristics, the ColorSTIP is expected to become an enhanced version of the STIP. As can be expected, because more information is being used to find the interest points, more points are detected, as result producing a denser set of features.

B. Proposal 2: HueSTIP

In this version, each space-time interest point is also represented in terms of the hue values quantized at the local spatiotemporal region of the point. The range of values for the hue is usually measured as angles in radians $(0 - 2\pi)$ and we divide this range into 36 bins to follow something similar to [15]. This lets each bin account for one range of hue values.

To construct the hue histogram, we calculate the bin number to which the hue value (of a pixel in the spatiotemporal volume) belongs with the formula $bin = hue * 36/2\pi$. Then, the saturation value at corresponding pixel is accumulated at the position bin to which the hue value was assigned. Before accumulating the amount of saturation in the histogram bin, the saturation is weighed by a corresponding value in a weighing Gaussian mask. This means that depending on the position of the pixel, the saturation will have a different contribution weight. For the pixels centered at the spatiotemporal volume, the saturation will have total participation when quantizing the hue histogram. The size and values forming the spatiotemporal

Gaussian mask will vary according to the spatial and temporal scales of the interest point.

By using the hue histograms to describe the spatiotemporal features, the color information should aggregate robustness to the extracted patterns in terms of illumination intensity, lighting geometry, viewpoints and specular reflection [15].

III. Experimental Setup

As for the experiments, we evaluated the performance of the proposals through a 5-fold cross validation scheme for the binary cases and test-train scheme for the multiclass one. The bag-of-features concept was applied to represent each video. Therefore, the histograms of features were constructed based on the low level features provided by the proposals. Training and testing were performed by means of SVM using the linear kernel for both the binary (violence and pornography) and multiclass (movie human actions) cases.

An outline of the steps taken to perfom the experiments is described below:

1) Extract local features of the whole dataset (using all descriptors, namely, STIP, HueSTIP, ColorSTIP and HueColorSTIP),
2) Create the dictionary of features (or visual words), one dictionary for each feature type (STIP, HueSTIP, ColorSTIP and HueColorSTIP),
3) Construct histograms of features for each video by counting the occurrence of visual words in its set of features;
4) Apply
   a) either cross-validation (for the binary cases, namely, pornography and violence):
      i) Separate the video dataset into 5 groups (which are called folds). Videos from each class are equally distributed over the set of folds;
      ii) Learn a classifier from the training set of each fold using SVM (one for each feature type). The training set of each fold are the features from the remainding folds (e.g., when learning the classifier for fold0, features from all other folds are used for training);
      iii) Test each fold for all classes contained in it.
   b) or the test-train scheme (for the multiclass case), where the training set is used to learn several binary classifiers by means of SVM (we used the libsvm [17]). The test samples are later evaluated by all learned classifiers to decide to which class each one of them belongs.

A. Movie Human Actions Dataset (Hollywood2)

We wanted to evaluate the performance of the descriptors for human action recognition in natural scenarios. Therefore, the Hollywood2 dataset [18] 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.

B. Pornography Database

The Pornography dataset attempts to portray the content diversity found on sharing social networks. For the pornographic class, videos were retrieved from websites specialized on the genre. The content varies in terms of the ethnicities (asians, blacks, whites, multi-ethnic) of the people appearing in the scenes, as well as with regard to the sexual orientation and sexual practices. We sampled 400 videos of the porn class, but for our experiments we have randomly selected 200 of them.

For the nonpornographic class, the video content comprised many different subjects: documentaries, educational videos, car races, TV programs, cartoons, music concerts, sport matches, dancing, interviews, daily news, among many others. Of the sampled 400 nonporn videos, 200 exhibit confounding characteristics related to the pornographic content, for example, women wearing bikinis at the beach and undressed babies being bathed. In those cases the exposure of skin imposes a challenge to the system. On the other hand, the other half of the nonporn set (more 200 videos) display contents totally visually unrelated to pornographic scenes. We also only selected 200 videos from the entire pornography dataset for our experiments. An illustration of the dataset is depicted in Figure 2.

C. Violence Database

The Violence dataset was one of the contributions of this work. We collected video clips from social networks specialized in fights (Figure 3). Both violence and nonviolence samples try to be diverse and representative. A compilation of daily life situations in schools, ghettos, entrance spaces of night clubs, matches from several sport variations ($e.g.$, soccer, hockey), traffic, involving both spontaneous fights and professional wrestling sports build the violence dataset. Scenarios are depicted by aggressive behaviors involving any number of people, in indoor and outdoor environments, with or without presence of moving objects in the background ($e.g.$, cars). The nonviolence dataset was created by copying from

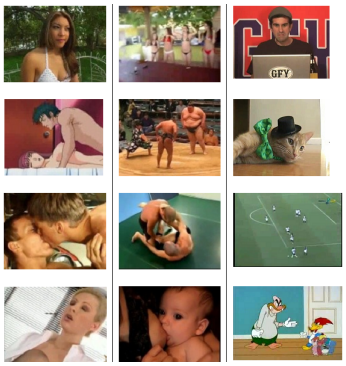

Fig. 2. Pornography dataset illustration. At the left, only videos of pornographic content. At the center, samples of difficult cases of nonpornographic content. At the right, we show the easy cases of nonpornographic content.

the negative class of the Pornography dataset videos that do not exhibit violent content. In total, there are 400 videos, 200 from each category.

Fig. 3. Violence dataset depicted by several scenarios of fights.

IV. Results and Discussion

The analysis of results is contextualized under different application contexts where one metric can be more important than others. In this regard, we empirically highlight which combination of interest point and descriptor can be more appropriate depending on the scenario requirements. As follows, we limit our discussion to some of the results available in the full text of the dissertation.

A. Human Actions Recognition: Video Annotation and Retrieval

[2]

By observing Table I, we can readily note that for a subset of the classes, all color-based STIPs were superior to the original STIP, namely *DriveCar*, *HandShake*, *Kiss*, *SitDown* and *StandUp*. For the *HandShake* and *Kiss* classes, increased rates can be explained by the scenes of close-up of hands and faces exhibited in the majority of videos of the classes (see Figures 4). However, it is important to keep in mind that motion features and shape information remain important to describe the human parts performing the actions. Therefore, color information is expected to enhance, but not to replace any of the other features.

[2]**We have published earlier results in [19]**

Fig. 4. Scenes of the *Kiss* and *HandShake* actions in which the objects of interest are focused by the camera.

Indoor environments (like offices, bedrooms, living rooms, dining rooms, kitchen) are what prevail in scenes of the *AnswerPhone*, *SitDown* and *Eat* classes. Characteristic colors of such scenarios can be helpful to contextualize the detected features defining the actions, which indeed happens to the cases of the *AnswerPhone* and *SitDown* classes. Interestingly, the *GetOutCar* class involves a very assorted scenario in terms of color, among other aspects, which makes it more confusing than helpful for representation of actions. Nevertheless, color information was apparently a decisive factor to improve performance by means of HueSTIP and Hue-ColorSTIP.

As expected, the skin color information taken from face close-ups in *Kiss* scenes were helpful to increase the number of hits by the ColorSTIP (see Figure 6). It is interesting to observe, however, that half of the mistakes made by the ColorSTIP for videos correctly classified by STIP were grayscale clips, as can be seen in Figure 5). Additionally, in one specific video the skin color is not shown at all (see Figure 5(G)).

TABLE I
This table reports the accuracy rates achieved by each version of the STIP on the Hollywood2 dataset (a multi-class dataset). The best overall performance for each version was chosen to show the results discriminated by classes.

Action	STIP	HueSTIP	ColorSTIP	Hue-ColorSTIP
AnswerPhone	9.4%	7.8%	**12.5%**	10.9%
DriveCar	71.6%	74.5%	**75.5%**	72.6%
Eat	**57.6%**	48.5%	33.3%	39.4%
FightPerson	64.3%	**67.1%**	**67.1%**	62.9%
GetOutCar	15.8%	**21.1%**	15.8%	19.3%
HandShake	6.7%	8.9%	13.3%	**15.6%**
HugPerson	**33.3%**	21.2%	21.2%	19.7%
Kiss	28.2%	37.9%	**50.5%**	45.6%
Run	**58.9%**	56.0%	56.0%	56.7%
SitDown	43.5%	45.4%	**50.9%**	48.2%
SitUp	**2.7%**	0.0%	**2.7%**	0.0%
StandUp	51.4%	61.0%	**63.0%**	58.2%
Average	36.9%	37.4%	**38.5%**	37.4%

The experimental results showed that there is indeed a gain

Fig. 5. Illustration of the set of *Kiss* action videos that were correctly classified by STIP, but not by COLORSTIP. Figures (B), (C), (D) and (G) are in gray scale, while the others are colorful. Figure (G) only depicts the sillhuoettes of the actors, so no skin color is shown.

Fig. 6. Illustration of the set of *Kiss* action videos that were correctly classified by COLORSTIP, but not by STIP

in the performance rates when using the color-based features to represent human actions patterns, for which the COLORSTIP won the race with the best overall performance.

B. Violence and Pornography Detection: Filtering of Unwanted/Prohibited Content

[3]

By observing the confusion matrices in Table II, we note that among the color-based algorithms COLORSTIP was the only one to fulfill the requirement of improving the true positive rates, accordingly reducing the rate of false negatives. For applications requiring the removal of abusive content, low rates of false negatives are more valuable. This is either because the interested audience does have strong moral reasons or because the target people are more sensitive. In other contexts, such as the case of forensics systems, false negatives are even more critical. It is desirable to extract the maximum number of suspect data (meaning the necessity of a higher recall), even if false positives are retrieved, provided that the average precision is high. False positives can be tolerated for applications of this nature, since they are less probable to contribute in either affecting a sensitive audience or invalidate forensics evidence. In this context, COLORSTIP shall be considered the best choice.

Note that in Case 1 of COLORSTIP mistakes (see Figure i of the supplemental material [4]), the actors are dressed during the whole film and in Case 2 only the woman's head is displayed in great part of the video. Other common error of ColorSTIP involved back story scenes in a big fraction of the videos. Most of those scenes do not exhibit sexual acts and nudity of the actors (see Figures iv and v). Regarding STIP, a great portion of its errors that differed from ColorSTIP's consisted of videos displaying clear scenes of explicit sex. Illustrations from Figures ii and iii of the supplemental material show visible examples of such mistakes. Another interesting case that we observed by analyzing all folds was that, unlike STIP, ColorSTIP hit all cases of pornographic content from cartoons. Also, it correctly detected as pornography photo-essays of porn stars, which can be possibly explained by the great amount of skin color presented in the scene. In such cases, the actions are usually softer, not well characterized by the fast-paced acts of sex.

We believe that the use of color for describing the motion patterns (HUESTIP) has slightly improved the results for the difficult nonpornographic cases. Yet, it weakened or turned ambiguous actual pornographic videos with visual aspect similar to the hard nonporn cases. However, interest points by color-dependent variations in space and time (COLORSTIP) gave a more arguably meaningful representation for the pornographic content. This was so expressive that it failed for one of the hard nonporn cases (Figure 7).

Fig. 7. Illustrative cases of COLORSTIP false positives, in a round where STIP made no mistakes with respect to the false positive cases.

TABLE II
CONFUSION MATRICES FOR THE PORNOGRAPHY CASE.

	STIP		**HUESTIP**		**COLORSTIP**	
Class	*Porn*	*NonPorn*	*Porn*	*NonPorn*	*Porn*	*NonPorn*
Porn	**.87**	.13	**.86**	.14	**.90**	.10
NonPorn	.075	**.925**	.06	**.94**	.08	**.92**

In the context of detecting violent scenes (see Table III),

[3] **We have reported and published results on those application contexts in [3], [4]**

[4] **As the content of the images may shock some sensibilities, we have made them available as a separate supplement. Available at http://www.npdi.dcc.ufmg.br/colorbasedstips/supplement.pdf**

however, STIP excels before its counterpart HUESTIP. This is not so surprising considering since we had already proven in previous work [4] that STIP was self-sufficient for distinguishing regular from aggressive events (which was one the contributions of the work). In that work its opponent was the state-of-the-art local feature detector SIFT, which was applied to frame samples of video shots instead of to the shots themselves. The STIP reached 100% of score, after the majority voting procedure over the visual histograms of the shots.

We believe that color information in violence scenes is not as clearly associated to specific actions as it apparently happens to the pornography's. Indeed, color information from videos depicting aggressive acts does not give any distinct characteristic for the type of actions. The people involved in the scenes are generally dressed, such that not much exposure of skin color is available. Furthermore, the background scenarios vary a lot in terms of color, which visually gives more muddling than deciding clues about the content.

TABLE III
CONFUSION MATRICES FOR THE VIOLENCE CASE

	STIP		**HUESTIP**	
Class	*Porn*	*NonPorn*	*Porn*	*NonPorn*
Porn	**.91**	.09	**.90**	.10
NonPorn	.15	**.85**	.155	**.845**

V. CONCLUSION

In this work, we have studied the impact of color-based motion patterns on the classification of actions. Color information was embedded in the STIP algorithm for detection and description of motion patterns. On this basis, we derived three color-based algorithms for extraction of spatiotemporal features (HUESTIP, COLORSTIP and HUE-COLORSTIP). The key idea was to enrich the spatiotemporal detector/descriptor with color information without losing important photometric invariance properties. For this purpose we have employed a well established color invariance model that is described in [15].

As Social Networks evolve, the need to provide tools for semantic classification and retrieval (including to control the proliferation of abusive content) becomes a critical issue. In addition, Digital Forensics has recently arisen as an exciting field of research, which can benefit from our tools to gather evidence from confiscated computers and hard disks, in cases of suspected child pornography, for example. In the latter case experts' time can be dramatically saved by preselecting among hundreds of thousands of documents those which should receive attention. Those and many other applications could take advantage of the proposed contributions.

ACKNOWLEDGMENT

The authors would like to thank the funding agencies CAPES, CNPq, FAPEMIG and FAPESP for the financial support to carry out this work. We are very grateful to Dr. Ivan Laptev for sharing with us the STIP source code, making this work possible.

REFERENCES

[1] N. T. Siebel and S. J. Maybank, "The advisor visual surveillance system," in *Proceedings of the ECCV 2004 workshop "Applications of Computer Vision" (ACV'04), Prague, Czech Republic*, M. Clabian, V. Smutny, and G. Stanke, Eds., May 2004, pp. 103–111.

[2] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005.

[3] F. D. M. de Souza, G. C. Chavez, E. A. do Valle Jr., and A. de A. Araujo, "Violence detection in video using spatio-temporal features," *Graphics, Patterns and Images, SIBGRAPI Conference on*, vol. 0, pp. 224–230, 2010.

[4] E. Valle, S. E. F. de Avila, A. da Luz Jr., F. D. M. de Souza, M. de M. Coelho, and A. de Albuquerque Araújo, "Content-based filtering for video sharing social networks," *CoRR*, vol. abs/1101.2427, 2011.

[5] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[6] W. Kelly, A. Donnellan, and D. Molloy, "Screening for objectionable images: A review of skin detection techniques," in *International Machine Vision and Image Processing Conference (IMVIP)*, 2008, pp. 151–158.

[7] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, and A. de A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1552–1556.

[8] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," in *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1*. Washington, DC, USA: IEEE Computer Society, 2002, p. 10433.

[9] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila, "Cassandra: Audio-video sensor fusion for aggression detection," in *IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS*, 2007.

[10] C. Clarin, J. Dionisio, M. Echavez, and P. Naval, "Dove: Detection of movie violence using motion intensity analysis on skin and blood," in *PCSC '06: Proceedings of the 6th Philippine Computing Science Congress*. Computing Society of the Philippines, 2006, pp. 150–156.

[11] J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," in *PCM '09: Proceedings of the 10th Pacific Rim Conference on Multimedia*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 930–935.

[12] T. Giannakopoulos, D. I. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *SETN*, 2006, pp. 502–507.

[13] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, 2003, pp. 432–439.

[14] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, vol. 32, no. 9, pp. 1582–1596, 2010. [Online]. Available: http://www.science.uva.nl/research/publications/2010/vandeSandeTPAMI2010

[15] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *ECCV*, vol. Part II. Springer, 2006, pp. 334–348. [Online]. Available: http://lear.inrialpes.fr/pubs/2006/VS06

[16] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[17] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[19] F. de Souza, E. Valle, G. C. Chávez, and A. de Albuquerque Araújo, "Color-aware local spatiotemporal features for action recognition," in *CIARP*, 2011, pp. 248–255.