Combining Orientation Tensors for Human Action Recognition

Virgínia F. Mota, Jéssica I. C. Souza, Arnaldo de A. Araújo
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{virginiaferm, jessicaione, arnaldo}@dcc.ufmg.br

Marcelo Bernardes Vieira
Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora
Juiz de Fora, Brazil
marcelo.bernardes@ice.ufjf.br

***Abstract*—This paper presents a new tensor motion descriptor based on histogram of oriented gradients. We model the temporal evolution of gradient distribution with orientation tensors in equally sized blocks throughout the video sequence. Subsequently, these blocks are concatenated to create the final descriptor. Using a SVM classifier, even without any bag-of-feature based approach, our method achieves recognition rates greater than those found by other HOG techniques on KTH dataset and a competitive recognition rate for UCF11 and Hollywood2 datasets.**

***Keywords*-Histogram of gradients; Orientation tensor; Motion description; Human action recognition.**

I. Introduction

Human action recognition is a research field with application in several areas such as video indexing, surveillance, human-computer interfaces, among others. Several works in the literature tackle this problem by extracting a set of descriptors and comparing them throughout a similarity measure. In the past years, several descriptors have been proposed. One of the features widely used to create descriptors for actions is the histogram of oriented gradients (HOG) [1]. In general, HOG is used as motion descriptor combined to other features or extracted around interest points [2].

The use of tensors is gaining space in image classification [3] and in video classification [4], [5]. An interesting work about tensor descriptors is presented in [6]. Perez et al [6] proposed to represent histogram of oriented gradients in orientation tensors in order to create a global motion descriptor. The drawback of this method is that the aggregation of several tensors could lead to an isotropic tensor, which does not have any direction information. Thus, we propose to improve this tensor descriptor. We argue that we could extract more information with a new tensor aggregation, which takes into account the individual motion information of each tensor.

Contributions: The main contribution of this work is to present a new motion descriptor based on orientation tensors. We model the temporal evolution of 3D gradient distribution with orientation tensors in equally sized blocks throughout the video sequence. Subsequently, these blocks are concatenated to create the final descriptor. Using a SVM classifier, even without any vocabulary computation, our method achieved recognition rates greater than those found by other HOG techniques on KTH dataset and a competitive recognition rate for UCF11 and Hollywood2 datasets.

A. Related work

Several descriptors used for human action classification consist in the combination of features. Laptev et al [7] proposed the combination of HOG with histogram of optic flow (HOF) to characterize local motion and appearance. These two features are computed and accumulated in space-time neighborhoods of detected interest points. Similarly to the SIFT descriptor [8], normalized histograms are concatenated to HOG and HOF vectors.

Two-dimensional features derived from histograms of oriented gradients have been shown to be effective for detecting human actions in videos. However, according to the viewing angle, local features may be often occluded. One alternative was proposed by Kläser et al [9] to avoid the problems presented by HOG2D. The authors presented the HOG3D, a descriptor based on histograms of 3D gradient orientations and can be seen as an extension of SIFT descriptor [8]. Gradients are computed using an integral video representation in spatio-temporal interest points detected by the Harris operator [10].

Recently, Wang at al [2] extended this idea of combination by modeling descriptors along dense trajectories. The time evolution of trajectories, HOG [1], HOF [1] and MBH (motion boundary histogram) [1], is modeled using a space time grid along trajectories.

Umakanthan et al [11] evaluated four popular descriptors (HOG, HOF, HOG/HOF, HOG3D) extracted around spatio-temporal interest points by Harris3D detector. They showed that the combination HOG/HOF performed better in all tested datasets.

In these four works, the signature for the whole video is computed using the popular bag-of-visual-feature method [12] and the descriptors are extracted in interest point neighborhood. As this method achieve the best results and is becoming very popular, there are few recent works that studies descriptors without using this technique.

Perez et al [6] proposed to combine HOG3D features into orientation tensors in order to create a global motion descriptor. It is important to note that histograms of oriented

gradients are similar to SIFT, but they are not computed only in salient points.

Chen et al [13] extracted a saliency map of each frame and accumulated them to form an Accumulative Edge Image (AEI). Subsequently, grid-based HOG is calculated on AEI to form a feature vector.

A global descriptor based on the histogram of oriented gradients is presented by [14], using the Weizmann database. The descriptor is computed using several time scales. From each scale, the gradient for each pixel is computed, resulting in a HOG for each video. Laptev et al [15] improved this descriptor using multiple temporal scales as the original and using multiple temporal and spatial scales.

The rest of the paper is organized as follows. In Section II, we present the fundamentals needed to a better understanding of our method. In Section III, we provide a detailed description of our approach. Finally, in Section IV, we carry out experiments on three benchmark action datasets.

II. Fundamentals

A. Orientation Tensor

An orientation tensor is a representation of local orientation which takes the form of an $m \times m$ real symmetric matrix for m-dimensional signals. Given a vector $\vec{v}$ with m elements, it can be represented by the tensor $T = \vec{v}\vec{v}^T$. Note that the well known structure tensor is a specific case of orientation tensor [16].

The geometric interpretation of this tensor is very attractive to motion description. Given a tensor of third order (a tensor in R^3), with eigenvalues λ_1, λ_2 and λ_3, it can be interpreted as following:

- $\lambda_1 \gg \lambda_2 \gg \lambda_3$ corresponds to an approximately linear tensor, with the spear component being dominant.
- $\lambda_1 \approx \lambda_2 \gg \lambda_3$ corresponds to an approximately planar tensor, with the plate component being dominant.
- $\lambda_1 \approx \lambda_2 \approx \lambda_3$ corresponds to an approximately isotropic tensor, with the ball component being dominant, and no main orientation present.

Thus, for motion description, we are interested in anisotropic tensors, which has direction information. This interpretation can be extended to m dimensions.

B. Histogram of gradients

The partial derivatives, or gradient, obtained by the filtering of the j-th video frame at pixel p is defined as the vector:

$$\vec{g}_t(p) = [dx\ dy\ dt] = \left[\frac{\partial I_j(p)}{\partial x}\ \frac{\partial I_j(p)}{\partial y}\ \frac{\partial I_j(p)}{\partial t}\right],$$

or, equivalently, in spherical coordinates $\vec{s}_t(p) = [\rho_p\ \ \theta_p\ \ \psi_p]$ with $\theta_p \in [0,\ \pi]$, $\psi_p \in [0,\ 2\pi)$ and $\rho_p = ||\vec{g}_t(p)||$. This gradient indicates brightness variation that might be the result of local motion.

The gradient of all n pixels of the frame I_j can be compactly represented by a tridimensional histogram of gradients $\vec{h}_j = \{h_{k,l}\}$, $k \in [1, nb_\theta]$ and $l \in [1, nb_\psi]$, where nb_θ and nb_ψ are the number of cells for θ and ψ coordinates, respectively. There are several methods for computing the HOG3D and we chose, for simplicity, a uniform subdivision of the angle intervals to populate the $nb_\theta \cdot nb_\psi$ bins:

$$h_{k,l} = \sum_p \rho_p \cdot w_p,$$

where $\{p \in I_j \mid k = 1 + \left\lfloor \frac{nb_\theta \cdot \theta_p}{\pi} \right\rfloor, l = 1 + \left\lfloor \frac{nb_\psi \cdot \psi_p}{2\pi} \right\rfloor\}$ are all points whose angles map to k and l bins, and w_p is a per pixel weighting factor, which can be uniform or Gaussian as in [8]. The whole gradient field is then represented by a vector $\vec{h}_j$ with $nb_\theta \cdot nb_\psi$ elements.

Different from [6], we do not need any extra parameters to reduce inter-frame brightness unbalance.

III. Proposed Method

A. Orientation tensor: coding HOG3D coefficients

Given the histogram $\vec{h}_f$ of frame f, the tensor of the frame can be defined as:

$$T_f = \vec{h}_f \vec{h}_f^T,$$

and carries the information of the gradient distribution of the frame f. Individually, this tensor has the same information of $\vec{h}_f$, but several tensors can be combined to find component covariances.

The temporal covariance of the gradient distribution can be given by $T = \sum_a^b T_f$ using all video frames or an interval of interest. By normalizing T with a L_2 norm, we are able to compare different video clips or snapshots regardless their length or resolution.

Since T is a symmetric matrix of m-order, it can be stored with $\frac{m(m+1)}{2}$ elements, where $m = nb_\theta \cdot nb_\psi$.

This series of orientation tensor can diverge, thus we obtain an isotropic tensor that does not hold useful motion information. However, if the series converge as an anisotropic tensor, it carries meaningful average motion information of the frame sequence. The conditions of divergence and convergence need further studies.

B. Tensor descriptor: analysis in blocks

When the gradient histogram is computed using the whole frame, the cells are populated with vectors regardless their position in the frame. This implies in a loss of the correlation between the gradient vectors and their neighbors. As observed by [8], the subdivision of the video into cubes of frames enhances the recognition rate, using a Gaussian weight for w_p.

Suppose the video frame f is uniformly subdivided in $\vec{x}$ and $\vec{y}$ directions by a grid with n_x and n_y non-overlapping blocks. Each block can be viewed as a distinct video varying in time. The smaller frames result in gradient histograms $\vec{h}_j^{i,j}$, $i \in [1, n_x]$ and $j \in [1, n_y]$, with better position correlation. The tensor for each block $b_{i,j}$ is then computed as the addition of all block tensors throughout the video:

$$T_{i,j} = \sum_f \vec{h}_f^{i,j}\ \vec{h}_f^{i,j\,T},$$

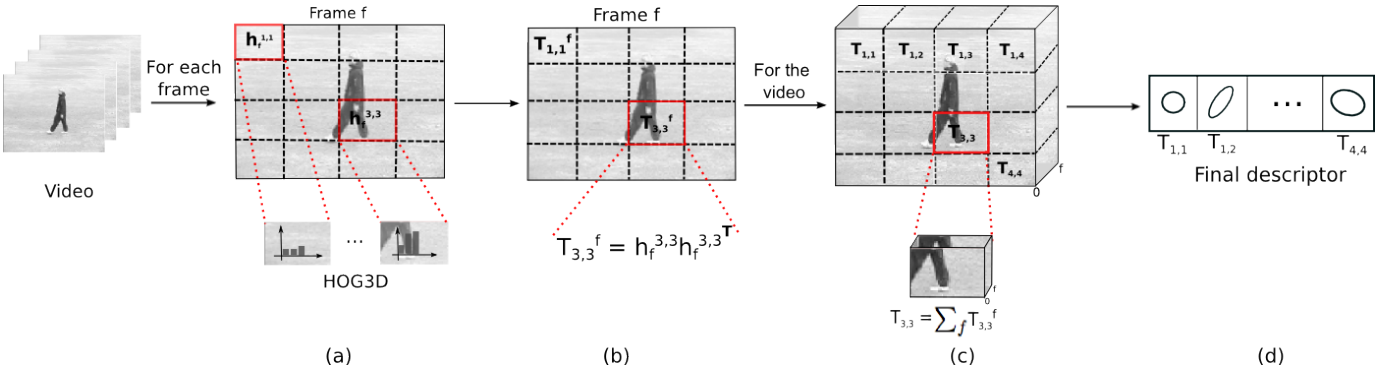

Fig. 1. Framework of the proposed combination of orientation tensors. This is an example of a tensor descriptor obtained from a 4x4 grid and each tensor block carries the information of the gradient distribution. The steps of our approach are: (a) Extract the 3D histogram of gradients for each subdivision of each video frame; (b) Code HOG3D into orientation tensors for each block; (c) Accumulate the frame tensor from each block in order to model the temporal evolution of gradients; and (d) Concatenate each tensor block.

which captures the uncertainty of the direction of the m-dimensional vectors $\vec{h}_f^{i,j}$ for each block.

Perez et al [6] lose this local covariance information when they aggregated all tensors in a unique series of orientation tensors. We argue that we could extract more motion information by analyzing each tensor block individually.

Thus, the final tensor descriptor is obtained by combining all blocks of the video. We propose to concatenate the individual block tensors, to form the final descriptor for the input video:

$$T = \{T_{i,j}\}_{1 \leq i \leq n_x, 1 \leq j \leq n_y}$$

where the size of the final descriptor depends on the number of bins and the number of subdivisions.

Figure 1 shows an example of a tensor descriptor obtained from a 4x4 grid. To a better understanding of the method, the tensors are represented as an ellipsis (that is a second-order tensor). In our case, all tensors are m-order, where m is the number of bins of the histogram ($nb_\theta \cdot nb_\psi$).

We could also improve the descriptor by accumulating the tensor obtained with the video frame flipped horizontally. The video frame is flipped and the rest of the framework remains the same. Then, this new information is simply added to the original tensor of the block. This flipped version enforces horizontal gradient symmetries that occur on the video, even those between multiple frames.

IV. Experimental Results

In our experiments, we used three benchmark video datasets: KTH [17], UCF11 [18] and Hollywood2 [19] datasets. We evaluated our descriptor in a classification task and followed the same evaluation protocol proposed by the authors of the datasets with a SVM classifier. For each dataset, we vary the number of subdivisions of the video (4x4, 8x8) in order to have a more local motion information. We also vary the number of bins of the HOG3D (2x4, 3x6, 4x8, 8x16) to have more orientations of the motion.

A. KTH Dataset

The KTH action dataset [17] consists of 6 human action classes. Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in most sequences. In total, the database consists of 2391 video samples (Figure 2).

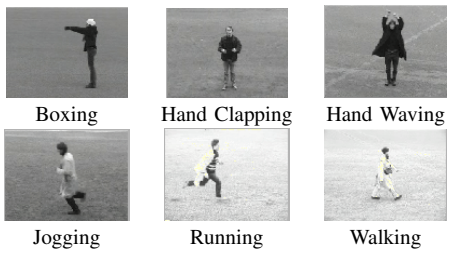

Fig. 2. Examples of videos from KTH dataset [17].

The performance of our method on the KTH dataset is reported in Table I. The confusion matrix for the best result is presented in Table II.

TABLE I
Recognition rate for KTH dataset for several parameter sets.

Parameters	Recognition Rate (%)
Grid 4x4 HOG 2x4	74.2
Grid 4x4 HOG 3x6	83.2
Grid 4x4 HOG 4x8	89.7
Grid 4x4 HOG 8x16	**92.5**
Grid 8x8 HOG 2x4	79.0
Grid 8x8 HOG 3x6	87.3
Grid 8x8 HOG 4x8	88.9

Perez et al [6] reported that the best result is 92.0% achieved with an 8x8 grid and a HOG 8x16. We can see that our new combination slightly improved the recognition rate, however, the mean difference of classification of each class for the best result is 0.48 $\pm$ 3.63, with 99.0% of confidence, that is, it is

TABLE II
CONFUSION MATRIX FOR KTH DATASET WITH A SUBDIVISION OF 4×4 AND A HOG WITH nb_θ=8 AND nb_ψ=16.

	Box	HClap	HWav	Jog	Run	Walk
Box	95.8	0.0	0.0	0.0	0.0	4.2
HClap	2.8	97.2	0.0	0.0	0.0	0.0
HWav	0.7	6.9	92.4	0.0	0.0	0.0
Jog	0.0	0.0	0.0	93.1	4.9	2.1
Run	0.0	0.0	0.0	18.1	81.9	0.0
Walk	0.0	0.0	0.0	5.6	0.0	94.4

not statistically significant. Indeed, the KTH dataset has only simple actions and this new combination does not extract much more information. This could happen because several tensors per block became isotropic, as they model parts of the video that do not contain significant motion. Nevertheless, the best result is achieved using less subdivisions.

B. UCF11 Dataset

The UCF11 dataset [18] (also known as UCF YouTube) consists in 11 action categories extracted from Youtube video sequences. The sequences can vary regard camera motion, viewpoint, background, illumination and object appearance, pose and scale. The dataset contains a total of 1168 sequences (Figure 3).

Fig. 3. Examples of videos from UCF11 dataset [18].

The performance of our method on the UCF11 dataset is reported in Table III. The classification per action class is presented in Table IV.

The results with this dataset are not reported in [6]. Thus, we evaluated the descriptor using the same parameters as ours. Table V presents recognition rates for several subdivisions for the descriptor proposed by [6]. We chose to use only HOG3D with 8x16 bins as they achieved the best results in [6]. The best recognition rate was 68.9% using a grid 32x32 and a HOG3D of 8x16 bins. In that way, we show that our descriptor is more robust than the descriptor previously described in [6].

The mean difference of classification of each class for the best result is 6.5 ± 4.9, with 99.0% of confidence. So, we can conclude that our descriptor improved significantly the recognition rate.

TABLE III
RECOGNITION RATE FOR UCF11 DATASET FOR SEVERAL PARAMETER SETS.

Parameters	Recognition Rate (%)
Grid 4x4 HOG 2x4	57.3
Grid 4x4 HOG 3x6	63.9
Grid 4x4 HOG 4x8	70.1
Grid 4x4 HOG 8x16	**75.4**
Grid 8x8 HOG 2x4	58.9
Grid 8x8 HOG 3x6	64.7
Grid 8x8 HOG 4x8	71.0

TABLE IV
AVERAGE PRECISION (AP) FOR EACH CLASS OF UCF11 DATASET WITH A SUBDIVISION OF 4×4 AND A HOG WITH nb_θ=8 AND nb_ψ=16.

Action	**AP(%)**
Biking	77.6
Diving	97.0
Golf	92.5
Juggle	50.0
Jumping	73.3
Riding	86.1
Shooting	61.2
Spiking	83.0
Swing	80.1
Tennis	68.9
WalkDog	59.8
Mean	**75.4%**

TABLE V
RECOGNITION RATES OF UCF11 DATASET FOR SEVERAL PARAMETER SETS FOR THE DESCRIPTOR PROPOSED BY [6].

Parameters	Recognition Rate (%)
Grid 4x4 HOG 8x16	65.0
Grid 8x8 HOG 8x16	67.5
Grid 16x16 HOG 8x16	68.4
Grid 32x32 HOG 8x16	**68.9**

C. Hollywood2 Dataset

The Hollywood2 dataset [19] consists of a collection of video clips extracted from 69 movies and categorized in 12 classes of human actions. In total, there are 1707 action samples divided into a training set (823 sequences) and a test set (884 sequences), where training and test samples are obtained from different movies (Figure 4).

The performance of our method on the Hollywood2 dataset is reported in Table VI. The average precision is shown in Table VII.

The best result reported by [6] is 34.0% achieved with a 4x4 grid and a HOG 8x16. It is interesting to note that

with the same parameters we could improve the result in 6%, which demonstrates that our descriptor carries more useful information than the descriptor previously proposed by [6].

The mean difference of classification of each class is 4.9 ± 4.8, with 99.0% of confidence. So, we can conclude that our descriptor improved significantly the recognition rate.

Fig. 4. Examples of videos from Hollywood2 dataset [19].

TABLE VI
RECOGNITION RATE FOR HOLLYWOOD2 DATASET FOR SEVERAL PARAMETER SETS.

Parameters	Recognition Rate (%)
Grid 4x4 HOG 2x4	23.4
Grid 4x4 HOG 3x6	31.2
Grid 4x4 HOG 4x8	34.5
Grid 4x4 HOG 8x16	**40.3**
Grid 8x8 HOG 2x4	25.2
Grid 8x8 HOG 3x6	33.0
Grid 8x8 HOG 4x8	36.7

TABLE VII
AVERAGE PRECISION (AP) FOR EACH CLASS OF HOLLYWOOD2 DATASET USING A SUBDIVISION OF 4×4 AND A HOG WITH nb_θ=8 AND nb_ψ=16.

Action	**AP(%)**
AnswerPhone	20.6
DriveCar	71.6
Eat	28.1
FightPerson	57.7
GetOutCar	20.3
HandShake	24.4
HugPerson	26.2
Kiss	55.2
Run	59.2
SitDown	54.4
SitUp	19.9
StandUp	45.5
Mean	**40.3%**

D. Comparison to the state-of-the-art

A comparison with the state-of-the-art methods is presented in Table VIII. We show published results reported in the literature. For all datasets, we reported the best recognition rate of our method which is achieved with a subdivision of 4×4 and a HOG with nb_θ=8 and nb_ψ=16.

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART FOR KTH, UCF11 AND HOLLYWOOD2 DATASETS.

KTH	
Method	Recognition Rate(%)
Laptev et al [15]	72.0
Mota et al [4]	86.4
Klaser et al [9]	91.4
Laptev et al [7]	91.8
Perez et al [6]	92.0
Umakanthan et al [11] (HOG3D)	92.6
Wang et al [2]	94.2
Our Method	92.5

UCF11	
Method	Recognition Rate(%)
Chen et al [13]	67.5
Perez et al [6]	68.9
Wang et al [2]	84.2
Our Method	75.4

Holywood2	
Method	Recognition Rate(%)
Perez et al [6]	34.0
Klaser et al [9], [20]	43.7
Laptev et al [7], [20]	45.2
Umakanthan et al [11] (HOG3D)	46.9
Wang et al [2]	58.3
Our Method	40.3

The proposed method achieved a competitive accuracy with a much simpler approach, using only the information of histograms of oriented gradients, without any bag-of-feature strategy [2], [7], [9], [11], that is, without the cost of a codebook computation.

In all datasets we improved the performance of the descriptors previously proposed by [6] and showed better results than other HOG based descriptors [7], [9], [13]. It is interesting to note that the recognition rates of our descriptor were close to the bag-of-feature approach evaluated in [11] using HOG3D and interest points. Moreover, they used a one-against-the-rest approach while we used a two-fold approach for classification.

Thereby, we can conclude that our descriptor aggregates useful information from HOG3D, enhancing the recognition rate. Moreover, it achieved promising results in all datasets without the cost of a bag-of-feature based method. The drawback of our method lies on the the descriptor size, which could become very large. One may select some tensors in order do reduce this problem. For example, an anisotropy measure could be used to rank the most interesting tensors.

V. CONCLUSION

In this work, we presented an improved tensor motion descriptor based on HOG3D. It is a simple and effective approach reaching 92.5% of recognition rate with KTH dataset, comparable to the state-of-the-art methods. However, for more challenging datasets, as the UCF11 and Hollywood2, we note that interest points information plays an important role

and bag-of-feature approaches improved overall recognition. Our recognition rate is lower than those obtained by these approaches but is fairly competitive in both datasets, even using only HOG information.

The main advantage of our method is that it reaches good recognition rates using only HOG information, which shows how promising is this technique. Moreover, it is possible to aggregate more information from other features in order to improve the results.

It is interesting to note the behavior of our method when we add more subdivisions. In some cases, increasing this number does not necessarily leads to a better classification, such as in UCF11 dataset. Indeed, the higher the number of subdivisions, the grater is the chance of the tensor block becomes isotropic, as it might represent a part of the video that does not have meaningful motion.

In order to improve the recognition rate of our descriptor, we intend to further analyze the combination of other video features. Furthermore, since the result improves when analyzing individual tensors, we need to study how our descriptor behaves with a bag-of-feature based method.

ACKNOWLEDGMENT

The authors would like to thank CNPq, CAPES and FAPEMIG for funding. This work uses RETIN SVM classifier [21].

REFERENCES

[1] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006. [Online]. Available: http://lear.inrialpes.fr/pubs/2006/DTS06

[2] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176.

[3] R. Negrel, D. Picard, and P. H. Gosselin, "Using spatial pyramids with compacted vlat for image categorization," in *ICPR*, 2012, pp. 2460–2463.

[4] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, and P.-H. Gosselin, "A tensor based on optical flow for global description of motion in videos," in *SIBGRAPI 2012 (XXV Conference on Graphics, Patterns and Images)*, Ouro Preto, MG, Brazil, august 2012, pp. 298–301.

[5] R. Negrel, V. Fernandes Mota, G. Philippe-Henri, M. Bernardes Vieira, and F. Precioso, "Indexation des Bases Vidéos à l'aide d'une Modélisation du Flot Optique par Bases de Polynômes," in *Actes de la conférence RFIA 2012*, Lyon, France, Jan. 2012, pp. ISBN : 978–2–9 539 515–2–3.

[6] E. A. Perez, V. F. Mota, L. M. Maciel, D. Sad, and M. B. Vieira, "Combining gradient histograms using orientation tensors for human action recognition," in *International Conference on Pattern Recognition*, 2012, pp. 3460–3463.

[7] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision & Pattern Recognition*, jun 2008. [Online]. Available: http://lear.inrialpes.fr/pubs/2008/LMSR08

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ser. ICCV '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523

[9] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, sep 2008, pp. 995–1004. [Online]. Available: http://lear.inrialpes.fr/pubs/2008/KMS08

[10] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, no. 2-3, pp. 107–123, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-1838-7

[11] S. Umakanthan, S. Denman, S. Sridharan, C. Fookes, and T. Wark, "Spatio temporal feature evaluation for action recognition." in *DICTA*. IEEE, 2012, pp. 1–8. [Online]. Available: http://dblp.uni-trier.de/db/conf/dicta/dicta2012.html#UmakanthanDSFW12

[12] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 1470–1477 vol.2.

[13] X. gan Chen, J. Liu, and H. Liu, "Will scene information help realistic action recognition?" in *Intelligent Control and Automation (WCICA), 2012 10th World Congress on*, 2012, pp. 4532–4535.

[14] L. Zelnik-manor and M. Irani, "Event-based analysis of video," in *In Proc. CVPR*, 2001, pp. 123–130.

[15] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Underst.*, vol. 108, pp. 207–229, December 2007.

[16] B. Johansson, G. Farnebck, and G. F. Ack, "A theoretical comparison of different orientation tensors," in *Symposium on Image Analysis*. SSAB, 2002, pp. 69–73.

[17] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.

[18] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

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

[20] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[21] J. Fournier, M. Cord, and S. Philipp-Foliguet., "Retin: A content-based image indexing and retrieval system," *Pattern Analysis and Applications*, vol. 4, pp. 153–173, 2001.