Action Recognition using the $\mathfrak{R}_\mathfrak{f}$ Transform on Optical Flow Images

Josep Maria Carmona and Joan Climent
Barcelona Tech (UPC), Barcelona, Spain
josep.maria.carmona@estudiant.upc.edu

Keywords: $\mathfrak{R}$ Transform, Action Recognition, PHOW, Projection Templates.

Abstract: The objective of this paper is the automatic recognition of human actions in video sequences. The use of spatio-temporal features for action recognition has become very popular in recent literature. Instead of extracting the spatio-temporal features from the raw video sequence, some authors propose to project the sequence to a single template first.

As a contribution we propose the use of several variants of the $\mathfrak{R}$ transform for projecting the image sequences to templates. The $\mathfrak{R}$ transform projects the whole sequence to a single image, retaining information concerning movement direction and magnitude. Spatio-temporal features are extracted from the template, they are combined using a bag of words paradigm, and finally fed to a SVM for action classification.

The method presented is shown to improve the state-of-art results on the standard Weizmann action dataset.

1 INTRODUCTION

One of the multiple applications of automatic visual analysis of human movements is the understanding of human activities in video sequences. Classification and recognition of human activities can be very useful for multiple applications, like video-surveillance, human-computer interaction, biometric analysis... The objective of action/gesture recognition is to identify human movements invariantly to the gesture speed, distance to camera, or background.

Since human activity is captured in video sequences, the temporal domain is very important to model gestures or human actions. Several authors have extended the classical object recognition techniques to the spatio-temporal domain (Kläser et al., 2008), (Scovanner, 2007), (Jhuang et al., 2007). They use vocabularies of volumetric features that are computed using three-dimensional keypoint detectors and descriptors. In (Scovanner, 2007) they used 3D SIFT for action recognition using spatio-temporal features. In (Kläser et al., 2008), authors proposed a descriptor based on the histogram of 3D spatio-temporal gradients. 3D gradients are binned into regular polyhedrons. They also extend the idea of integral images to 3D which allows rapid dense sampling of the cuboid over multiple scales and locations in both space and time. The approach presented in (Jhuang et al., 2007) combined keypoint detection with the calculation of local descriptors in a feed-forward framework. This was motivated by similarity with the human visual system, extending a bioinspired method to action recognition. At the lowest level, they compute the spatial gradients along the *x* and *y* axis for each frame. Then the obtained responses are converted to a higher level using stored prototypes.

We present a new method for action/gesture recognition, based on a projection template, which is obtained using a variant of the $\mathfrak{R}$ transform. The $\mathfrak{R}$ transform was originally designed for object recognition, but some authors (Souvenir and Parrigan, 2009) (Wang et al., 2007) (Zhu et al., 2009) (Vishwakarma et al., 2015) (Goudelis et al., 2013) have used it, or some variants, for action recognition too. In (Vishwakarma et al., 2015), they proposed a method based on the combined information obtained from the $\mathfrak{R}$ transform and the energy silhouettes. They generated a feature vector from the average energy silhouettes and applied $\mathfrak{R}$ transform to the normalized silhouette extracted. Authors presented in(Goudelis et al., 2013) two methods to assess the capability of the Trace transform to recognize human actions. Trace transform is a generalization of the Radon transform.

These previous works use this transform on silhouette images or human shapes previously segmented from image sequences. In this work, we apply the $\mathfrak{R}$ transform directly to the optical flow components of the input sequence, avoiding all the problems regarding the segmentation stage.

We also show in this paper that some variants of

Carmona J. and Climent J.
Action Recognition using the Rf Transform on Optical Flow Images.
DOI: 10.5220/0006218002660271
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 266-271
ISBN: 978-989-758-225-7

the $\mathfrak{R}$ transform preserve important information of human action sequences, giving more accurate results in the recognition process.

We compute different $\mathfrak{R}$ transforms using different projection functions. Using a $\mathfrak{R}_f$ transform, being f a projection function, we obtain a single image from each video sequence. Different projection functions f lead to different templates. In the results section we evaluate several projection functions and select the one that gives the highest recognition rate.

Once the image template is computed, we use a Pyramid Histogram Of visual Words (PHOW) (Bosch et al., 2007) as feature descriptor. Next, we combine the feature descriptors, ignoring the structural information among keypoints, using the paradigm known as Bag_of_Words (BoW) (Csurka et al., 2004). In a BoW approach, the number of occurrences of similar feature patterns is accumulated in the bins of a histogram. Some other authors have successfully used BoW for action recognition (Niebles et al., 2008).

Once the feature patterns have been computed, the action sequence is recognized by means of a SVM classifier. In the results section, we compare the results obtained using our approach on the Weizmann action dataset with the ones reported by other authors using the same dataset.

2 PRELIMINARIES

The Radon transform (Radon, 1917) consists in a multiple angle projection of a given image $I(x,y)$. The result of this projection is an integral line, that is, the cumulative sum of pixel values in all directions. Given a line in its polar form:

$$\rho = x cos\theta + y sin\theta \quad (1)$$

the Radon transform can be expressed mathematically using equation 2

$$g(\rho,\theta) = \sum_x \sum_y I(x,y)\delta(x cos\theta + y sin\theta - \rho) \quad (2)$$

where $I(x,y)$ is the input image, δ is the Dirac function, ρ is the distance from the line to the origin, and θ is the line direction. The main drawback of the Radon transform is that it is not invariant to translation, scale, or rotation. There exist several approaches to achieve such invariances (Arodz, 2005). In (Tabbone et al., 2006), they presented a variant of the Radon transform, the $\mathfrak{R}$ transform, which is invariant to translation and scale.

The $\mathfrak{R}$ transform is computed summing all squared values of the Radon transform for all image rows having a given direction θ. It can be expressed using equation 3

$$\mathfrak{R}(\theta) = \sum_\rho g^2(\rho,\theta) \quad (3)$$

The result of the $\mathfrak{R}$ transform is a function giving the normalized sum of pixel values for all orientations. It maps a 2D image to a 1D signal.

The $\mathfrak{R}_f$ transform is a variant of the $\mathfrak{R}$ transform, being f a generic function. It can be expressed in its general form:

$$\mathfrak{R}_f(\theta) = f(g(\rho,\theta)) \quad (4)$$

where $g(\rho,\ \theta)$ is the Radon transform and f is a function that can be tuned as parameter, and allows to adapt the transform to the specific problem to be solved.

For example, $\mathfrak{R}_{max}$ substitutes the squared values of the $\mathfrak{R}$ transform by the maximum of the absolute value of pixel values. This transform is invariant to translation and, if correctly normalized dividing by the supremum of the image values, is also invariant to scale.

$$\mathfrak{R}_{max}(\theta) = \left\{ \begin{array}{ll} \underset{\rho}{max}(g(\rho,\theta)) & if \quad R_1 \geq R_2 \\ \underset{\rho}{min}(g(\rho,\theta)) & if \quad R_1 < R_2 \end{array} \right\} \quad (5)$$

where $R_1 = |\underset{\rho}{max}(g(\rho,\theta))|$ and $R_2 = |\underset{\rho}{min}(g(\rho,\theta))|$.

$\mathfrak{R}_{dev}$ uses the standard deviation instead the sum of squared values. It is also invariant to translation.

$$\mathfrak{R}_{dev}(\theta) = \underset{\rho}{dev}(g(\rho,\theta)) \quad (6)$$

$\mathfrak{R}_{mean}$ uses the man value of pixel values for all orientations. Even though is pretty similar to the original $\mathfrak{R}$ transform, it has the advantage of considering the negative values of $g(\rho,\theta)$.

$$\mathfrak{R}_{mean}(\theta) = \underset{\rho}{mean}(g(\rho,\theta)) \quad (7)$$

The properties of all these $\mathfrak{R}_f$ transform are totally dependent on the function f chosen. Apart from their properties concerning invariances, they present different behaviors when applied to images that may contain negative values (like the optical flow images used in this work). Figure 1 shows the result of the $\mathfrak{R}$, $\mathfrak{R}_{max}$, $\mathfrak{R}_{dev}$, $\mathfrak{R}_{mean}$ for two input images containing positive and negative values. We can see that $\mathfrak{R}_{max}$ and $\mathfrak{R}_{mean}$ transform give different results for positive and negative values, while these differences are lost when using the standard $\mathfrak{R}$ and the $\mathfrak{R}_{dev}$ transforms.

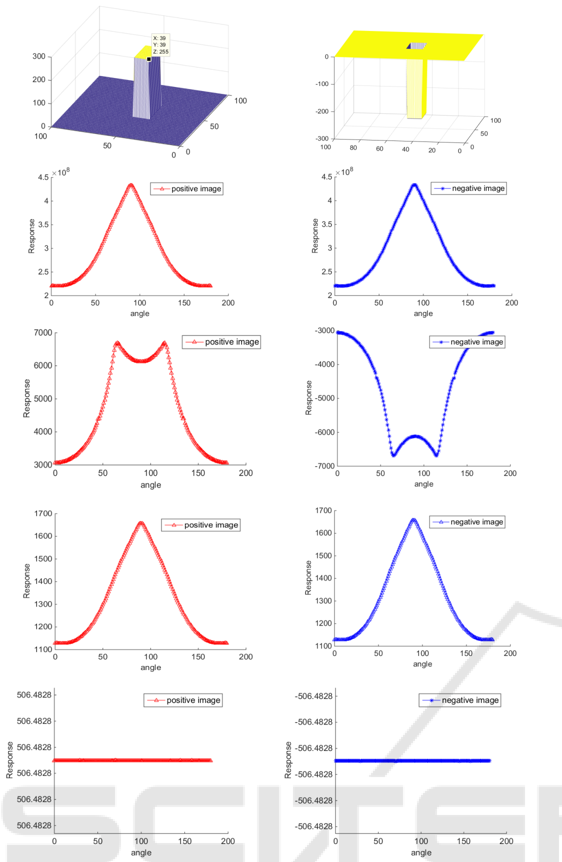

Figure 1: Result of the $\mathfrak{R}$, $\mathfrak{R}_{max}$, $\mathfrak{R}_{dev}$ and $\mathfrak{R}_{mean}$ respectively, for two sintetic images containing positive (left column) and negative (right column) values.

In this work, we consider to use the $\mathfrak{R}$ transform, and its variants, on human action video sequences, but instead of applying them to grey level, shape, or edge images like in (Souvenir and Parrigan, 2009), (Wang et al., 2007) and (Zhu et al., 2009), we do it on their optical flow components.

3 OUR APPROACH

This section describes the method used for action recognition in this work. It is based on a Bag of features approach, but previous to the keypoint extraction stage, the video sequences are projected to static templates. Figure 2 shows a block diagram of the whole process. Next, we describe in detail the different stages of our system. The implementation details, including the tuning parameters, are given in section 4.

In order to project the video sequences, we apply the $\mathfrak{R}_f$ transform to both F_x, F_y components of the optical flow, obtaining two surfaces $\mathfrak{R}_{fx}$ and $\mathfrak{R}_{fy}$.

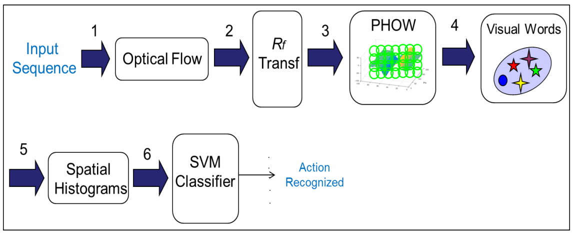

Figure 2: Block diagram of our approach.

These surfaces can be considered as spatio-temporal templates defining an action sequence. These surfaces retain some information of the different speeds that action movements have in a local region of the scene. Figure 3 shows an example of such surfaces, concretely the result of applying the $\mathfrak{R}_{max}$ transform on the 'bend' image from the Weizmann action dataset (Blank et al., 2005).

The optical flow of the video sequence has been computed using the real-time algorithm presented in (Karlsson and Bigun, 2012).

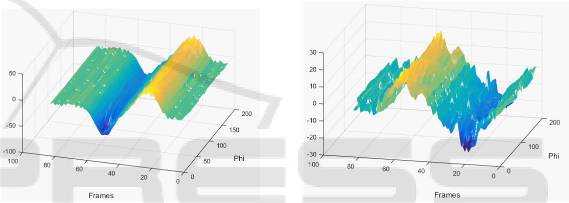

Figure 3: $\mathfrak{R}_{fx}$ and $\mathfrak{R}_{fy}$ surfaces computed applying $\mathfrak{R}_{max}$ transform on the 'bend' image from the Weizmann action dataset.

Once the transform has been computed, we search a set of keypoints in both surfaces using a standard detector. In this work we have used the PHOW detector (Bosch et al., 2007), a variant of SIFT but computed on a dense grid at different scales. Figure 4 shows the keypoints obtained using PHOW on the images shown in figure 3. Only 50 keypoints, randomly chosen, are shown for visualization purposes. The circles are centered on the selected keypoints, their sizes represent the scale, and the lines inside the circles show the main gradient orientation. Once the keypoints have been selected, we use a classical descriptor based on a gradient orientation vector. We use a 128 bins histogram for each keypoint to describe the gradient orientations within a local neighborhood.

Using a Bag of features approach, similar descriptors are grouped using a k-means clustering technique. The set of cluster centers form a visual codebook. In this way, every input sequence is represented by a set of words, each word describing a small region of the $\mathfrak{R}_f$ surfaces. For the classification stage, we use a Stochastic Dual Coordinate Ascent Methods (SDCA)(Shalev-Shwartz and Zhang, 2013) linear

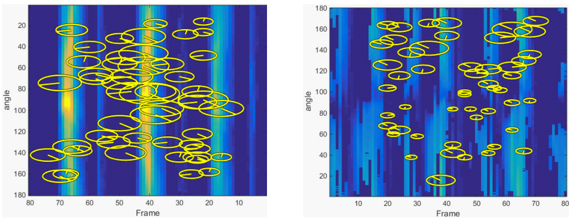

Figure 4: Keypoints obtained using PHOW on the $\mathfrak{R}_{fx}$ and $\mathfrak{R}_{fy}$ surfaces.

SVM solver. We tested several kernels and we obtained the best performance using Chi-Squared Kernel (χ^2) (Vedaldi and Zisserman, 2012). It can be expressed using equation 8.

$$k(x,y) = \sum_{i=1}^{n} \frac{2x_i y_i}{(x_i + y_i)} \tag{8}$$

Where x, y are the n-element input vectors.

4 RESULTS

In our preliminary experiments we have used the Weizmann dataset (Blank et al., 2005). It is a widely used sequence database containing a set of human actions. The sequences have been recorded with static camera and background, there are no occlusions, and only a person is moving in all sequences. They do not present serious illumination changes either. This dataset consists in 10 different actions carried out by 9 different persons. Figure 5 shows some snapshots of the Weizmann dataset.

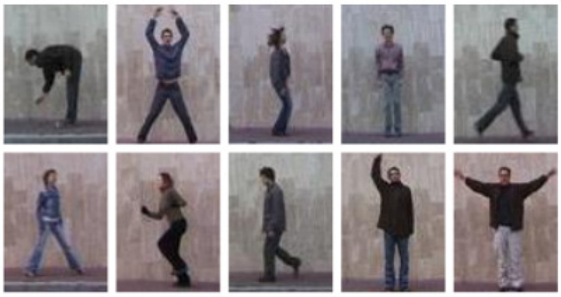

Figure 5: Weizmann human actions. Bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2.

For the experiments, we have used leave-one-out cross validation method, since it is the usual method used by other authors for testing. We use the 10 actions done by a single person for testing, and the actions done by the remaining 8 persons are used for training. This process is repeated for all 9 persons.

We have done two different experiments. The first one has the objective of determining which one of the $\mathfrak{R}_f$ transforms will produce a higher action recognition rate. As evaluation criteria we have computed the recognition rate ($\overline{RR}$) using all four different transforms ($\mathfrak{R}$, $\mathfrak{R}_{max}$, $\mathfrak{R}_{mean}$, and $\mathfrak{R}_{dev}$).

$$\overline{RR} = \frac{\text{samples correctly classified}}{\text{Total samples Tested}} \tag{9}$$

Once we have determined the optimum $\mathfrak{R}$ transform, it is selected for the second experiment. The objective of this second experiment is to compare the results of our approach with the ones published by other authors using the same dataset.

For tuning the parameters of our approach, we have tested codewords from 100 to 1100 visual words. For PHOW we have tested from 1 to 5 pixels between keypoints in the grid of dense SIFT and from 2 to 10 for the size of the spatial bins (scales). We have obtained the best performance using 1-pixel distance between keypoints in the grid of dense SIFT, a single scale of size 3, and a Visual Vocabulary of 900 visual words.

Table 1 shows the results obtained for the four different transforms $\mathfrak{R}_f$($\mathfrak{R}$, $\mathfrak{R}_{max}$, $\mathfrak{R}_{mean}$, and $\mathfrak{R}_{dev}$) using the scheme shown in the figure 2. We can see that using the supremum as projection function ($\mathfrak{R}_{max}$) we obtain the best results.

Table 1: Recognition rate (%) for $\mathfrak{R}$ and $\mathfrak{R}_f$ transforms applied to Weizmann dataset.

Transform	%
$\mathfrak{R}$	95.55
$\mathfrak{R}_{max}$	**98.88**
$\mathfrak{R}_{mean}$	88.55
$\mathfrak{R}_{dev}$	95.55

To establish a fair comparison with the state-of-art methods, we have tuned all parameters exactly as reported by the authors in their original papers. For the Scovanner method (Scovanner, 2007) we have used the configurations of the sub-histograms 2x2x2 and 4x4x4, and 8x4 histograms to represent θ and φ in the 3D SIFT descriptor. For the Kläser method (Kläser et al., 2008) we have used a code-book size V=4000, spatial and temporal support s0=8, t0=6, amount of histogram cells M=4, N=3, number of supporting mean gradients S=3, cut-off value c=0.25, and a complete orientation polyhedron icosahedron. For the Jhuang method (Jhuang et al., 2007) we used 500 gradient-based features. For the Niebles method (Niebles et al., 2008) the parameters chosen were σ =1.2, τ =1.2, the descriptors dimensionality was 100, and the codebook size was fixed to 1200. For (Vishwakarma et al., 2015) we have used a feature vector of [1 * 7 + 1 * 168 + 1 * 2] dimensions. For the method presented in (Goudelis et al., 2013) we have used Linear Discriminant Analysis (LDA) and a vector of 31 features. This latter method requires a previous silhouette extraction stage.

Table 2 shows the comparative results for all these methods using the Weizmann sequences. In our approach we have used the $\mathfrak{R}_{max}$ transform since it was the one that gave higher accuracy in the former experiment.

Table 2: Recognition rate (%) on the Weizmann dataset.

Method	%
(Scovanner, 2007)	84.2
(Kläser et al., 2008)	84.3
(Niebles et al., 2008)	90
(Jhuang et al., 2007)	98.8
(Vishwakarma et al., 2015)	96.64
(Goudelis et al., 2013)	93.4
Ours (using R_{max})	**98.8**

The mean computational time for recognizing an action sequence of 100 frames of 160x120 pixels is 900ms. It has been computed using a 3.1GHZ i3 Intel Core. For these preliminary tests, code is not optimized and the whole process has been implemented using Matlab.

5 CONCLUSIONS

Template based approaches allow to project a whole sequence into a single image. In this paper we have presented a generalized form $\mathfrak{R}_f$ of the Radon transform for projecting the action sequence. Choosing the correct projection function f, it can be adapted to a concrete problem.

We have tested three different f functions to project the Radon transform, namely, mean, standard deviation and supremum, and applied these transforms to the optical flow components of a video sequence. This experiment has shown that the $\mathfrak{R}_{max}$ transform gives the highest recognition rate for action recognition, higher than the standard $\mathfrak{R}$ transform, and the other projection functions.

The results obtained in a second experiment also show that the use of such transforms is a very promising technique since it yielded higher recognition rates than the state-of-art methods using the same dataset, achieving a 98.8 % recognition rate.

6 FURTHER WORK

The results presented in this paper have been obtained using the Weizmann dataset as testbed. Next experiments will involve other popular action/gesture recognition datasets like KTH (Schuldt et al., 2004) dataset, and Cambridge hand-gesture data set.

We are currently working on the extension of this technique to action segmentation. Standard action recognition datasets used by most researches, usually contain a set of single actions that start at the beginning of the sequence and stop at the end. In a real application actions should be detected, segmented, and finally, recognized. The use of the $\mathfrak{R}_f$ transforms is a promising technique for sequence segmentation too.

ACKNOWLEDGEMENTS

This work was supported by the Spanish Ministry of Science and Innovation, project DPI2016-78957-R, and AEROARMS European project H2020-ICT-2014-1-644271

REFERENCES

Arodz, T. (2005). Invariant object recognition using radon-based transform. *Computers and Artificial Intelligence*, 24:183–199.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. *Computer Vision, IEEE International Conference on*, 2:1395–1402 Vol. 2.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classification using random forests and ferns. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. A. (2004). Visual categorization with bags of keypoints. pages 1–22.

Goudelis, G., Karpouzis, K., and Kollias, S. (2013). Exploring trace transform for robust human action recognition. *Pattern Recognition*, 46(12):3238 – 3248.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8.

Karlsson, S. and Bigun, J. (2012). Lip-motion events analysis and lip segmentation using optical flow. pages 138–145.

Kläser, A., Marszaek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *In BMVC08*.

Niebles, J., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Akad. Wiss.*, 69:262–277.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA. IEEE Computer Society.

Scovanner (2007). A 3-dimensional sift descriptor and its application to action recognition. pages 357–360.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599.

Souvenir, R. and Parrigan, K. (2009). Viewpoint manifolds for action recognition. *J. Image Video Process.*, 2009:1:1–1:1.

Tabbone, S., Wendling, L., and Salmon, J.-P. (2006). A new shape descriptor defined on the radon transform. *Comput. Vis. Image Underst.*, 102(1):42–51.

Vedaldi, A. and Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492.

Vishwakarma, D., Dhiman, A., Maheshwari, R., and Kapoor, R. (2015). Human motion analysis by fusion of silhouette orientation and shape features. *Procedia Computer Science*, 57:438 – 447.

Wang, Y., Huang, K., and Tan, T. (2007). Human activity recognition based on r transform. In *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Zhu, P., Hu, W., Li, L., and Wei, Q. (2009). *Human Activity Recognition Based on R Transform and Fourier Mellin Transform*, pages 631–640. Springer Berlin Heidelberg, Berlin, Heidelberg.