An evaluation of the grid geometry for human action recognition

Pau Agustí[a,c], V. Javier Traver[a,c], Raúl Montoliu[b,c] and Filiberto Pla[a,c]

pagusti@lsi.uji.es vtraver@lsi.uji.es montoliu@icc.uji.es pla@lsi.uji.es

[a]Dept. of Computer Languages and Systems
[b]Dept. of Computer Science and Engineering
[c]iNIT (Institute of New Imaging Technologies)
University Jaume I, Castellón, Spain

Abstract

Visual recognition of human action in video sequences is, nowadays, one of the most emerging fields in computer vision. Most of the existing methods extract the features from a similar grid geometry of the bounding box covering the actor performing the action without taking into account new possibilities which can improve the recognition. This paper evaluates and compares a new polar geometry with the traditional cartesian one. Experimental results show that the proposed geometry improves the recognition rate significantly.

1 Introduction

At present, the analysis of video sequences in which people act has become one of the most emerging investigation fields in computer vision. That new challenge is due to its high impact on large technical and social applications: surveillance, human-machine interfaces, automatic diagnostics of orthopedic patients, analysis and optimization of athletes' performances, etc. This new area, where it is tried to recognize people acting, is commonly called human action recognition. In particular, we are interested in full-body human action recognition.

Some good surveys detail a large number of techniques for characterizing, learning and recognizing human actions and give taxonomies as [12, 13]. Others like [15] discuss different classification methods for human actions.

In human action recognition, it is extremely important to choose wisely the visual features to characterize and discriminate actions. Many possibilities have been considered most of them using the components of the optical flow for the descriptors [5, 4] globally within the interest region. However, local histograms and optical flow are quite recurrent in the human action recognition investigations [1, 3, 14, 11, 9, 8]. One important issue when using local histograms is how they are computed and which are their local image support regions. Typically rectangular grids have been considered [11, 9]. Alternative grids such as circular have scarcely been addressed [14]. In contrast, the main contribution of this paper is the consideration of alternatives for the locations where features are extracted from. Two alternative grid geometries (cartesian and polar) used for extracting features are considered in order to evaluate which one improve the recognition of human actions.

The rest of the paper is organized as follows. In the next section, the proposed method is explained, how the region of interest is detected and how the local histograms are computed. Then, some experiments are done, in the section 3, in order to evaluate which grid geometry is better for recognition. Finally, a conclusion is made and how the further investigation will be carried out.

2 Human action recognition

The complete method (which is summarized in Figure 1) for the creation of the feature vector, can be divided into three phases. In the first one, the subject performing the action is enclosed and this region of interest is prepared for its later processing; afterwards optical flow between two correlative frames is calculated, and orientation local histograms are updated from the optical flow following locations in the geometry grid.

Only the part of the frame where the person is located contains relevant information for the recognition. Therefore, all the background pixels must be discarded. For this reason, all computations can be focused on the person. After the frame from the sequence is obtained it is segmented and, by applying a connected components algorithm, the bounding box enclosing the subject (BB) is found. After that, a new subimage is created from the BB_t and it is resized, to a fixed predefined size ($W \times H$), in order to have all the new frames at the same size for the optical flow calculation.

Next, the optical flow, F_t, between two consecutive resized bounding boxes BB_t and BB_{t-1} is computed. We use the classical Lucas-Kanade's algorithm [10] (L&K) that provides a good dense estimation with a reasonable computational cost. In recent years, new optical flow algorithms have been proposed [2] which can be very accurate, but at the cost of an increased implementation complexity. However, the use of a more accurate optical flow method do not necessarily increase the recognition rate, as shown in [11].

Finally, the features are extracted. There are a lot of possibilities when creating the local histograms. For example, the orientation histogram could be weighed by the magnitude as in [9, 14], or using the optical flow cartesian components and the silhouette as in [1]. But otherwise, in this work we have been inspired by [11], and use orientation histograms. The contribution of an optical flow orientation to a histogram is weighed by its magnitude and the relative position of the pixel at which the optical flow is computed. The choice is based on the relevance of the orientation in contrast with the magnitude. Each bounding box, BB_t, is divided into N regions and an orientation histogram is created for each one. Since this is one of the key aspect of this work, it will be explained at length below along with the main contribution (the influence of the grid geometry for human action recognition).

2.1 Defining the feature vector

Given the bounding box BB_t with their optical flow calculation, F_t, already done, the BB_t is divide into N regions. For the moment we do not care about the form of the region cause the proposed method is valid for any grid geometry. In order to create the feature vector, the following steps are performed:

1. For each one of the N regions, a local orientation histogram which contains the information accumulated along time is created. Each histogram has n_θ bins corresponding to n_θ possible orientations of the optical flow in the range $[0 - 2\pi]$.

2. Four contribution function are also created at the beginning of the process. These are one-dimensional functions of the position z which are defined as $w(z) = w(z; K, L)$ where K is the number of positions (for example, pixels) and L the number of divisions. Thus, $\frac{K}{L}$ would be the number of positions in a single region. The first two functions, w_1 and w_2, are two piece-wise linear function, like can be seen in Figure 2.

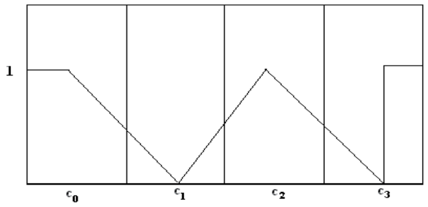

Figure 2: Weighing function, w_1, for a grid with $L = 4$ regions

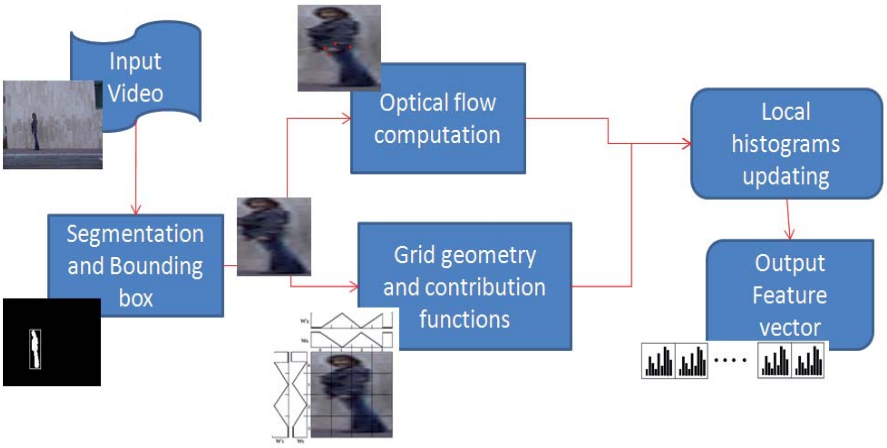
Input Video
Segmentation and Bounding box
Optical flow computation
Grid geometry and contribution functions
Local histograms updating
Output Feature vector

Figure 1: Overview of the method for the feature vector creation

The other two functions are the contribution to the neighboring region in one of the dimensions, w'_1, and the contribution to the neighbor region in the other dimension, w'_2. The values of those two functions are simply:

$$w'_1(z) = 1 - w_1(z)$$
$$w'_2(z) = 1 - w_2(z)$$

These contribution functions allow overlapping between the regions. This means that a pixel not only contributes to the region where it is included but also to the neighboring ones. Other functions rather than linear and other overlapping strategies could be defined.

3. Next, each pixel (x, y) in the bounding box BB_t is processed in order to update the histograms which it must contribute to. This process consists of calculating the magnitude $m(v_x, v_y)$ and orientation $\theta = \theta(v_x, v_y)$ of the optical flow at (x, y). Then, the region where this pixel lies on $R(x, y)$ and its neighboring regions along dimension i, $R_i(x, y), i \in 1, 2$ are identified. Finally, the histograms are updated like this:

$$H_{R(x,y)}[n_\theta(\theta)] \leftarrow H_{R(x,y)}[n_\theta(\theta))] + \\ + m(v_x, v_y) \cdot w_1(x) \cdot w_2(y)$$

$$H_{R_1(x,y)}[n_\theta(\theta))] \leftarrow H_{R_1(x,y)}[n_\theta(\theta))] + \\ + m(v_x, v_y) \cdot w'_1(x) \cdot w_2(y)$$

$$H_{R_2(x,y)}[n_\theta(\theta))] \leftarrow H_{R_2(x,y)}[n_\theta(\theta))] + \\ + m(v_x, v_y) \cdot w_1(x) \cdot w'_2(y)$$

where the $n_\theta(v_x, v_y)$ means the bin where the orientation, θ, is included.

For example, and to summarize, in the cartesian domain, one of the dimensions is x and the other is y. In the example shown in Figure 3, a rectangular grid it is used but is only to show clearly the contribution of a pixel in the different regions, and for the polar grid the same idea is applied. The pixel p is in the region $(1, 1)$ which is $R(x, y)$ and the contribution to that region is $w_x \cdot w_y$. The neighboring region in the x dimension is $(1, 0)$ which is $R_1(x, y)$, because it is the closest to the central pixel of that region, and the contribution $w'_x \cdot w_y$. Finally, the neighboring region in the y dimension is $(0, 1)$ which is $R_2(x, y)$, for the same reason, and the contribution $w_x \cdot w'_y$

4. Once every bounding box, BB_t, in the sequence are computed, the N histograms

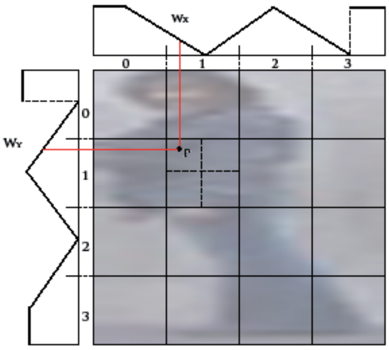

Figure 3: Rectangular grid showing the contribution functions for the pixel p

are normalized, and they are concatenated to create the feature vector:

$$\left(\frac{H_1}{h}, ..., \frac{H_N}{h}\right)$$

where $h = \sum_{i=1}^{N} \sum_{j=1}^{n_\theta} H_i[j]$ is the normalizing constant.

2.2 Partitioning the bounding box

The process described above is general in the sense that the particular geometry of the grid used to define the regions is not relevant for the method. However we think that grid geometry may play an important role in human action recognition. Most previous work using local histograms have considered rectangular grids [11, 9], and only few works have used a circular (but not polar) subdivision [14]. We propose a simple evaluation of two grid geometries but extending the circular one because the polar division is more biological inspired and it seems intuitively that it can be a good partitioning geometry. For instance, in a different but related problem [6] the polar partitioning is shown to improve the distinction of human from robots using their motion.

The first way to create a grid of regions is using the classical rectangular form (Figure 4, left). We define the number of regions that we want as $N = m \times n$ where m is the number of rows and n the number of columns. For the contribution functions we use cartesian coordinates, so the four functions for this geometry are: $w_1 = w_x = w(x; W, m)$, $w_2 = w_y = w(y; H, n)$, $w'_1 = w'_x = 1 - w_x$ and $w'_2 = w'_y = 1 - w_y$, where W is the weight and H the height of the BB.

The second way is using a polar grid (Figure 4, right). Here the central regions have less pixels but the same importance, because they have their own histogram, than the peripheral regions. We define the number of regions that we want as $N = s \times r$ where s is the number of sectors and r the number of rings. Finally the size or the positions in the radius is denoted by R. For the contribution functions we use polar coordinates, so the four functions for this geometry are: $w_1 = w_\rho = w(\rho; R, r)$, $w_2 = w_\phi = w(\phi; 2\pi, s)$, $w'_1 = w'_\rho = 1 - w_\rho$ and $w'_2 = w'_\phi = w_\phi$. Notice, that in this case, the contributions will be between neighbor regions in the radius direction (ρ) and in the angular direction (ϕ).

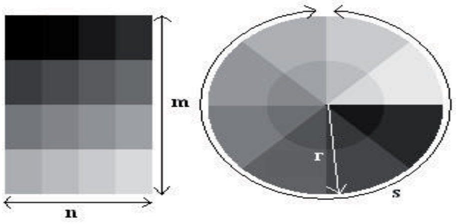

Figure 4: Rectangular ($m = 4$, $n = 4$, $N = 16$) and polar ($r = 2$, $s = 8$, $N = 16$) grids

2.3 Classifying actions

Once all the feature vectors are created we use the simple discriminative classifier K-nearest neighbors. The choice of this algorithm is because it is well known that it is very simple and very efficient. A leave-one-out cross-validation technique is used, training with all the actors except one and testing only with the one out. This estimation method is an unbiased assessment of the probability of success of the classifier, but with high variance [16].

3 Experiments

As it has already been explained most past works have focused on which kind of features are more discriminative for the human action recognition. However, the central point of this work is to explore how the geometry of the grid affect to the recognition rate. Normally, it is used a rectangular grid and extract characteristics from each rectangle in the grid. We compare that proposal with a polar grid where the central information is more relevant in comparison to the size in pixels (less pixels but the same importance in terms of the feature vector).

For the experiments the Weizman dataset [7] has been used. Two different experiments have been carried out with action sets of different complexities. For the two experiments the same parameters have been used, and they are detailed in Table 1. The segmentation of the human figure is taken from the masks available in the dataset. The first experiment has been performed using a subset of 4 out of the 9 Weizman dataset's actions. The 4 actions are: run, walk, waving with 1 hand (wave1) and waving with 2 hands (wave2). The choice of those 4 actions are so that two of them are similar between them and very different to the other two. Another different aspect between these two groups is that the actions in one group (run and walk) involves global body translation but not the actions in the other group. For this experiment 4 different distributions of the regions: 4×4, 8×2 (or 2×8), 2×12 (or 12×2) and 8×3 (or 3×8) have been compared.

Table 1: Parameters for the experiments

Parameter	Symbol	Value
Number of bins	n_θ	8
Size of the resized BB_t	$W \times H$	60×60
Window size in L&K	-	5×5

First, it will be explained the relevance of the distributions of the regions into the grid separately the cartesian and the polar geometries. Table 2 shows that the rectangular regions has better recognition rate by increasing the number of rows than the columns. The reason of that could be because the body is symmetric in the horizontal direction but not in the vertical direction having in this direction more information. The polar grid improves the recognition rate by increasing the number of divisions of the sectors and not by increasing the number of divisions of the radius.

Second, the comparison of the performances of both geometries is analyzed. In general, it can be seen that the recognition rate is higher with the polar grid than with the rectangular grid. A possible explanation of this is because the polar geometry fits better the structure of the human figure and its limbs than the rectangular grid.

Table 3: Confusion matrix (results in %) for the rectangular grid (2×12)

wave1	88.0	12.0	0.0	0.0
wave2	0.0	100.0	0.0	0.0
run	0.0	12.5	75.0	12.5
walk	0.0	0.0	43.0	57.0
	wave1	wave2	run	walk

Table 4: Confusion matrix (results in %) for the polar grid (12×2)

wave1	100.0	0.0	0.0	0.0
wave2	0.0	100.0	0.0	0.0
run	0.0	0.0	75.0	25.0
walk	0.0	0.0	0.0	100.0
	wave1	wave2	run	walk

Confusion matrices for the best results with the rectangular and polar grids are shown in Table 3 and 4, respectively. In the polar grid

Table 2: Classification results in % for the different configurations

	Number of features						
	128			192			
Grid type	4 × 4	8 × 2	2 × 8	12 × 2	2 × 12	3 × 8	8 × 3
Rectangular	73	74	66	74	80	70	77,5
Polar	80	90	70	93	70	75	90

and for this experiment only the run action is confused, this is normal because if someone does not run fast could be even confused by a human observer. In the rectangular grid and for this experiment more confusions are made. The worst is that run is sometimes confused with wave2 this is because someone moves excessively the arms and this movement is considered in the same regions than the wave2.

For the second experiment the whole action set is used but only with the 12×2 distribution for the polar grid and the 2×12 distribution for the rectangular grid. The global recognition results can be shown in Table 5. As could be observed using the whole action set the recognition rate is also better when the grid has a polar geometry than when has a rectangular geometry.

In the Tables 6 and 7 the confusion matrix of the results for the rectangular geometry (2 × 12) and for the polar geometry (12 × 2) using the whole action set are shown respectively. The comparison of both tables show that, in all the actions except one (wave1), the recognition rate of the polar geometry outperforms that of the rectangular one.

Table 5: Classification results in % for whole action set experiments

	Rectangular	Polar
Grid size	2 × 12	12 × 2
Recognition rate	65	73

4 Conclusion and further research

Human action recognition is one of the most emerging fields in computer vision due to its relevant applications for the daily life. Large amount of research has been done, but partitioning of the region of interest has not been evaluated before. For that reason, in this work the classical rectangular grid geometry is compared with a polar grid geometry. The results reveal that the classical grid geometry may be the simplest, but not the best performing choice. The polar grid fits better both the human structure and the motion of its limbs, which play an important role in human actions.

In further research this polar grid will be evaluated with more databases in order to confirm the preliminary conclusions. It could also be interesting to evaluate different possibilities of many elements defining the method such as the contribution functions as well as improving the action recognition rate.

5 Acknowledgements

The authors acknowledge the Spanish research programme Consolider Ingenio-2010 CSD2007-00018.

References

[1] M. Ahmad and Seong-Whan Lee. HMM-based human action recognition using multiview image sequences. *18th International Conference on Pattern Recognition, 2006. ICPR 2006.*, 1:263 –266, 2006.

Table 6: Confusion matrix (results in %) for the rectangular grid (2 × 12)

bend	55.0	23.0	11.0	0	0	11.0	0	0	0	0
jack	0	89.0	0	0	0	0	0	0	0	11.0
jump	0	0	78.0	0	11.0	0	11.0	0	0	0
pjump	50.0	12.5	0	25.0	0	0	0	12.5	0	0
run	0	0	0	0	80.0	0	10.0	10.0	0	0
side	11.0	0	0	0	0	89.0	0	0	0	0
skip	0	0	12.5	0	50.0	0	25.0	12.5	0	0
walk	0	0	0	0	0	20.0	0	80.0	0	0
wave1	23.0	0	0	11.0	0	0	0	0	55.0	11.0
wave2	11.0	22.0	0	0	0	0	0	0	0	67.0
	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2

Table 7: Confusion matrix (results in %) for the polar grid (12 × 2)

bend	67.0	0	11.0	11.0	0	11.0	0	0	0	0
jack	0	100.0	0	0	0	0	0	0	0	0
jump	0	0	78.0	0	11.0	0	11.0	0	0	0
pjump	25.0	12.5	0	50.0	0	0	0	12.5	0	0
run	0	0	0	0	89.0	0	0	11.0	0	0
side	11.0	0	0	0	0	89.0	0	0	0	0
skip	0	0	0	0	67.0	0	33.0	0	0	0
walk	0	0	0	0	0	0	0	100.0	0	0
wave1	44.5	0	0	0	0	0	0	0	44.5	11.0
wave2	11.0	11.0	0	0	0	0	0	0	0	78.0
	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2

[2] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV 2004. LNCS*, volume 3024, pages 25–36. Springer, Heidelberg, 2004.

[3] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *In European Conference on Computer Vision*. Springer, 2006.

[4] Somayeh Danafar and Niloofar Gheissari. Action recognition for surveillance applications using optical flow and SVM. In *ACCV (2)*, pages 457–466, 2007.

[5] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, Nice, France, 2003.

[6] Dario Figueira, Plinio Moreno, Alexandre Bernardino, José Gaspar, and José Santos-Victor. Optical flow based detection in mixed human robot environments. In *ISVC '09: Proceedings of the 5th International Symposium on Advances in Vi-*

sual Computing, pages 223–232, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[8] Nazli Ikizler, Ramazan Gokberk Cinbis, and Pinar Duygulu. Human action recognition with line and flow histograms. In *ICPR*, pages 1–4, 2008.

[9] X. Li. HMM based action recognition using oriented histograms of optical flow field. *Electronics Letters*, 43(10):560 – 561, 10 2007.

[10] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[11] Manuel Lucena, Nicolás Pérez Blanca, José Manuel Fuertes, and Manuel Jesús Marín-Jiménez. Human action recognition using optical flow accumulated local histograms. In *IbPRIA: Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, pages 32–39. Springer-Verlag, 2009.

[12] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

[13] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, November 2006.

[14] Filiberto Pla, Pedro Ribeiro, José Santos-victor, and Re Bernardino. Extracting motion features for visual human activity representation. In *IbPRIA: Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, 2005.

[15] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, September 2008.

[16] A. Webb. *Statistical Pattern Recognition*. Wiley, 2002.