Robust visual tracking combining global and local appearance models

Zhisheng Xu\(^1\), Shengping Zhang\(^1\), Julong Pan\(^2,3\), Xin Sun\(^1\) and Shaohui Liu\(^1\)

\(^1\)School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
\(^2\)College of Computer Science, Zhejiang University, Hangzhou, 310027, China
\(^3\)Department of Computer, China Jiliang University, Hangzhou, 310018, China

\{zsxu,spzhang,xsun,shaohl\}@vilab.hit.edu.cn julongpan@gmail.com

ABSTRACT

In this paper, we present a robust visual tracking method combining global and local appearance models. We model the object to be tracked with a RGB color histogram and multiple histograms of oriented gradients (HOG). Modeling object using only the former, a global appearance model, is widely used in visual tracking. However, it suffers many challenges such as illumination changes and pose changes and so on. In order to overcome this problem, we also model the object with multiple block based HOG histograms. The HOG histogram is a local appearance model and can effectively represent the shape information of the object which also gain increasing interests in computer vision especially in pedestrian detection. These two appearance models are complementary and used in the particle filter tracking framework. We test the performance of the proposed method on several challenging sequences, which verifies that our method outperforms the standard particle filter and achieves significant improvement.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: General; D.2.7 [Software Engineering]: Distribution and Maintenance—documentation

General Terms

Theory

Keywords

Visual tracking, particle filters, histograms of oriented gradients, global and local appearance models

1. INTRODUCTION

Visual tracking is one the most important subjects in computer vision with a wide range of applications — some of which are activity analysis, classification and recognition from motion and human-computer interfaces. The main goal of visual tracking is to imitate the motion sensibility of physical visual system, empower the machine with the ability of perceiving the object motion and their relations in the scene and provide an important way for image sequence understanding. Particle filtering is a widely used visual object tracking framework that is highly extensible and offers the flexibility to handle non-linearity and non-normality in the object models. In recent years, many new particle filter-based approaches have been proposed to solve difficult object tracking problems. However, most of this work has little much attention on how to combine the global and local appearance of the target to be tracked.

In this work, we present a tracking algorithm that considers both global and local appearance models. The target is modeled by two kinds of histograms corresponding to global and local appearance models, respectively. The global histogram is a RGB color histogram which is extracted in the whole region of the target. The local appearance model are multiple histograms of oriented gradients (HOG). The national behind this idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge direction. It captures edge or gradient structure that is very characteristic of local shape, so it has an easily controllable degree of invariance to local geometric and photometric transformations. The target is partly represented by multiple HOG histograms of multiple rectangular sub-regions (blocks) of the target. By measuring HOG histogram similarities between blocks of the target frame and the reference model, we got likelihoods for each block and then combine the likelihoods in a robust manner. With this manner, the spatial information of the target is effectively represented by dividing the target into multiple blocks. Finally we combine the similarities of the global RGB histogram and local HOG histograms to obtain the final observation likelihood in the particle filter framework.

2. RELATED WORK

2.1 Particle Filter

Particle filtering is a Monte Carlo approximation to the optimal Bayesian filter \([2]\), which monitors the posterior probability of a first-order Markov process through the following approximation formula:

\[
p(x_t | y_{1:t}) \approx \alpha p(y_t | x_t) \sum_{i=1}^{N} \pi_{t-i}^{(i)} p(x_t^{(i)} | x_{t-1}^{(i)}) \tag{1}
\]
Here, \(x_t\) is the process state at time \(t\), \(y_t\) is the observation, \(y_{1:t}\) is all of the observations through time \(t\), \(p(y_t|x_t)\) is the observation likelihood distribution, and \(\alpha\) is a normalizing factor. Each \(x_t^{(i)}\) is an instantiation of the process state, known as a particle, and the \(\pi_t^{(i)}\)'s are the corresponding particle weights. For each time \(t\), \(\sum_{i=1}^{n} \pi_t^{(i)} = 1\).

The sample set is updated by comparing each sample to the template. Each particle of the set is then reweighted in terms of the observation. At each time step, given the previous particle set \(\{x_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}\), a basic sequential importance resampling particle filter updates the particles as follows:

- Sample \(n\) particles \(x_{t-1}^{(i)}\) with replacement from current particle set according to probabilities \(\pi_{t-1}^{(i)}\).
- Generate an updated particle set by sampling from the proposal distribution, \(x_t^{(i)} \sim q(x_t^{(i)}|x_{t-1}^{(i)}, y_t, t)\).
- Reweight each particle according to the following formula and normalize so that the \(\pi_t^{(i)}\) sum to 1:

\[
\pi_t^{(i)} \propto \frac{p(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)}, y_t, t)} \tag{2}
\]

The mean state of an object is estimated at each time step by \(\bar{x}_t = \sum_{i=1}^{n} \pi_t^{(i)} x_t^{(i)}\). Particle filtering provides a robust tracking framework for it models uncertainty. Less likely object states have a small probability in the whole particle set. It keeps its options varied and consider multiple state hypotheses at the same time. There are many variants of particle filter for visual tracking [5, 6, 9, 4] proposed in the literatures.

### 2.2 Histogram of oriented gradients

Histogram of oriented gradient (HOG) [3] descriptors are feature descriptors used in computer vision and image processing for the purpose of object detection [7]. The technique counts occurrences of gradient orientation in localized portions of an image. This method is on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. The Histogram of Oriented Gradients descriptor maintains a few key advantages over other descriptor methods. Because the Histogram of Oriented Gradients descriptor operates on localized cells, the method upholds invariance to geometric and photometric transformations such changes would only appear in larger spatial regions. Coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HOG descriptor is thus particularly suited for human detection in images.

#### 2.2.1 Gradient computation

The first step in many feature detectors is to ensure normalized color and gamma values. However, this step can be ignored in HOG descriptor computation as the ensuing descriptor normalization essentially achieves the same result. Instead, the first step is the computation of the gradient values. Simply applying the 1-D centered, point discrete derivative mask in one of or both the horizontal and vertical directions is the most common method.

#### 2.2.2 Orientation binning

The second step of calculation involves creating the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation. The histogram channels are evenly spread over 0 to 180 degrees or 0 to 360 degrees, depending on whether the gradient is "unsigned" or "signed". Dalal and Triggs found that unsigned gradients used involved with 9 histogram channels performed best in their human detection experiments. In the experiments, the gradient magnitude itself generally produces the best results for the vote weight.

#### 2.2.3 Descriptor blocks

The gradient strengths must be locally normalized in order to account for changes in illumination and contrast, which requires grouping the cells together into larger, spatially-connected blocks. The HOG descriptor is then the vector of the components of the normalized cell histograms from all of the block regions. There are two main block geometries: rectangular R-HOG blocks and circular C-HOG blocks.

#### 2.2.4 Block normalization

There are a few methods for block normalization. Dalal and Triggs explored four methods all four methods showed very significant improvement over the non-normalized data.

#### 2.2.5 SVM classifier

Dalal and Triggs use Histogram of Oriented Gradient for human recognition. Once trained on images containing some particular object, the SVM classifier can make decisions regarding the presence of an object, such as a human being, in addition test images.

### 3. OVERVIEW OF THE METHOD

The study shows that locally normalized histogram of oriented gradient provide excellent performance relative to other existing feature sets. We calculate histogram of oriented gradients as follows: First we take the simplest scheme to compute the gradient for each pixel. For a pixel \((x, y)\), the gradient \(G(x, y)\) and magnitude \(M(x, y)\) of this pixel are computed as

\[
G(x, y) = \arctan\left(\frac{y(x, y+1) - y(x, y-1)}{x(x+1, y) - x(x-1, y)}\right) \tag{3}
\]

\[
M(x, y)^2 = (y(x, y+1) - y(x, y-1))^2 + (x(x+1, y) - x(x-1, y))^2 \tag{4}
\]

for color images, we calculate separate gradients for each RGB color channel, and take the largest norm one as pixel’s gradient vector.

Then each pixel calculates a weighted vote for an edge orientation histogram channel based on the orientation of the gradient element centered on it. The votes are accumulated into orientation bins over local spatial regions called “cell”. The orientation bins are evenly spaced over \(0^\circ - 360^\circ\). The vote is the magnitude of the vector. And each pixel in the block is weighted discriminatingly according to the distance to the center of the block.

The next step is to normalize the cells in one block. Gradient strengths vary over a wide range owing to local variations in illumination and foreground-background contrast, so effective local contrast normalization turns out to be essential
for good performance. The final descriptor is then the vector of all components of the normalized cell responses from all of the blocks in the detection window.

In order to handle partial occlusions or pose change and take into account the spatial distribution of the pixel intensities—information which is lost in traditional histogram-based algorithms, we used a robust statistic method in order to combine the vote maps of the multiple blocks.

We compared the histogram of oriented gradients of the particle to the template in each block, got the dissimilarities between blocks. Then we change it into a likelihood map by the follow formula, for each block \(i:\)

\[
L_g(i) = K \exp^{-\alpha v(i)}
\]

\(L_g(i)\) is the likelihood of \(i^{th}\) block in the region using gradient of orient gradients. \(v(i)\) is the dissimilarities of the block with the template we sorted the likelihood map and choose only the 40 percent blocks with larger probability for we assume that at least 2/5 blocks' visibility can guarantee the visible of the target. Then the likelihood is the average of the likelihood of chosen blocks. The fragments-based tracking [1] is effective when people was covered by other objects.

The method of using only the histogram of oriented gradients only concerns local appearance while global appearance was abandoned. We put forward a method concern both global and local appearance by calculating the likelihood between particles and template in

\[
L = \alpha \cdot L_g + (1-\alpha) \cdot L_c
\]

\(L_g\) stands for the likelihood by using histogram of oriented gradients, \(L_c\) stands for the likelihood by using histogram of RGB color, and \(L\) is the finally likelihood. We experiments on several videos. We compared the result of the tracker with the hand-labeled ground truth data in terms of relative distance and find \(\alpha = 0.5\) to have the better performance than \(\alpha = 0\) or \(\alpha = 1\). Proving that combining global and local appearance together has better performance than only considering global or local appearance.

4. EXPERIMENT RESULTS

In order to verify the effectiveness of the proposed method, we conduct experiments on several challenging video sequences which involve severe pose changes and fast moving. The proposed method is compared with the state-of-the-art algorithms: standard particle filter tracker. The comparison is done based on qualitative evaluation by looking at tracked results provided by the algorithms, and quantitative evaluation in term of tracking error [8]. The relative distance is defined as the normalized distance

\[
\text{Relative Distance} = \sqrt{\left(\frac{x-x_0}{s_x}\right)^2 + \left(\frac{y-y_0}{s_y}\right)^2}
\]

where \((x,y)\) is the location of the tracked target, \((x_0,y_0)\) and \((s_x,s_y)\) are the location and size of the ground truth. A perfect tracking expects the relative distance to be around 0.

We picked three representative of them here, the red line is the case when \(\alpha = 0\), the green line is the case \(\alpha = 0.5\) and the blue line is when \(\alpha = 1\).

Fig. 1 is the result of a hockey video, we track one of the players. From the result we can see that, compared with the method only considering global or local appearance, our method is more stable and is easy to reduce inaccuracy. In some cases, the result of using the histogram of oriented gradients or the histogram of RGB color alone is bad, and combining them has better performance. The quantitative evaluation results are shown in Fig. 2.

Fig. 3 is the result of a American football record, we track one of the players. In this case, the color of the sports clothes of the football player is similar to the boundary. So the method only considering global appearance may not work well. However, taking the local appearance into account can effectively reduce the inaccuracy. The quantitative evaluation results are shown in Fig. 4.

Fig. 5 is the result of a toy car video. We track the car. In this case, the car sometimes may be partly covered by other objects, causing it unstable to the histogram of RGB color. However, in calculating the likelihood of the histogram of oriented gradients, we only use the 40 percent with large probability. So the cover has little influence on the local appearance. The quantitative evaluation results are shown in Fig. 6 which also verify that the proposed method out-performs others.

5. CONCLUSIONS

![Figure 1](image1.png)  
**Figure 1:** The qualitative evaluation results of the hockey sequence for frames #0, #59, #115. The first row are the results obtained when \(\alpha = 0\). The second row are the results obtained when \(\alpha = 0.5\). The third row are the results obtained when \(\alpha = 1\).

![Figure 2](image2.png)  
**Figure 2:** The quantitative evaluation results of the hockey sequence.
In this paper, we propose a novel and effective visual tracking algorithm which combines the global and local appearance models in the particle filter framework. The proposed tracking method is more robust to illumination changes and pose changes since it effectively models the target’s color information and shape information together. In the proposed framework, the standard particle filter is the special case when the probabilistic fusion weight is set to 0. Experimental results on three challenging test sequences show that the proposed method outperforms standard particle filter tracking method.

6. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.61071180 and No. 60775024), National Basic Research Program of China (No. 2009CB320906) and Space Science Foundation (No. 20105577015).

7. REFERENCES