Realtime Object Matching with Robust Dominant Orientation Templates

Chaoqun Hong
College of Computer Science, Xiamen University of Technology
cqhong@xmut.edu.cn
Jianke Zhu, Mingli Song, Yinting Wang
College of Computer Science, Zhejiang University
jkzhu, brooksong, wangyt@zju.edu.cn

Abstract

Most of conventional object matching methods are based on comparing the local features, which are too computational demanding to achieve realtime performance on object detection in videos. Recently, Dominant Orientation Templates (DOT) method was proposed to make online feature detection and comparison feasible. However, it still suffers the problem of fragility due to the noise and partial occlusions. To efficiently tackle these problems, we introduce the similarity map to store the matching scores of individual grids in each sliding window, which is used to further denoise and infer the occlusion map. The promising experimental results demonstrate that proposed approach improves the robustness, which outperforms DOT assuredly.

1. Introduction

Object matching in large image collections and videos is now the burning issue in the computer vision research [8]. Differently from the conventional content-based image matching, it aims at not only finding the images related to the query object, but also providing the actual location information.

Being capable of dealing with different objects [5] using very simple scheme, template matching has been proven to be promising for object matching. The key idea of template matching is to combine the naive brute-force sliding window searching with the fast matching. Early approaches to template matching are dependent on contours and employ Chamfer distance as the similarity measure [2]. The major drawback of these methods is that the performance is greatly affected by the fragile contour extracting methods using the binary thresholding results. Some advanced template matching methods are presented by taking advantage of the image gradients [9]. One of the most successful features is Histograms of Gradients (HOG) [1], which captures the local distributions of image gradients computed on a regular grid. However, extracting the HOG features involves with the high computational cost. Moreover, it is not easy to adapt it to the realtime applications. Ouyang et al. [6] computed the rectangle sums and orthogonal Haar transform (OHT) based on integral images, which is more efficient than the naive brute-force searching methods. Some rotation-invariant features, i.e., Rotation-Invariant Fast Features (RIFF) [10] and Fourier Coefficients of radial projections (Forapro) [4], achieve the good matching performance.

Most recently, Dominant Orientation Templates (DOT) reveal the promising direction [3]. Differently from HOG feature, it quantizes the orientations of gradients by grid and only keeps the most significant orientations. Benefited from the compact feature representation, it utilizes the bitwise operations and a partial updating scheme to improve the speed. Therefore, DOT achieves realtime performance on a regular PC.

However, there are still some problems. First, it is fragile to noise since the scores of grids indicate whether two grids are similar rather than how similar they are. Second, the performance of DOT declines when partial occlusion occurs. Because it employs a naive scheme to compute the global score within the whole sliding window. It improves the speed of template matching while discards the local scores.

Generally speaking, there are always a lot of objects in the real-world scenes and the layout of them is very complicated. Moreover, the occlusion among the objects commonly occurs in the object matching task, which severely deteriorates the performance of object matching systems. Therefore, in this paper, we introduce a novel robust DOT matching approach. The con-
tribution is two-fold:

- We construct DOT similarity map which consists of the scores of each grid and reflects the visibility of the object.
- We use median filter to remove noise and connected component analysis to get the maximum visible part of the target.

2 Occlusion Handling with DOT Similarity Maps

The proposed occlusion handling approach is inspired by the idea of HOG-LBP [11], in which Wang et al. constructed an occlusion likelihood map for each scanning window by utilizing the response of each block of the feature to the global detector. As this map indicates the visibility of each block, it is then segmented by Mean-Shift clustering algorithm to estimate the occluded regions and the un-occluded regions. The part detectors are applied to obtain the final classification. This method depends on the classification scores of SVM, which is very computational expensive.

In contrast to HOG-LBP, the proposed occlusion handling approach makes use of the energy function of DOT to reduce the computational cost. There is a flaw in the original implementation of energy function. The energy function of DOT performs AND operation on 128 bits in parallel and counts the total number of 1’s using two other operations. Therefore, the similarities of 16 grids are computed with the lookup table simultaneously. It improves the speed of the matching process, however, the local scores are lost and only a global decision will be made. Although we can loose the detecting threshold to make DOT tolerant to some partial occlusions, it may increase the chances of false detections. Thus, we have to adapt it to improve the performance in the case of the partial occlusions.

Based on the above idea of occlusion handling, we try to modify the results of lookup tables of DOT. Instead of directly calculating the total number of similar grids by adding the similarity scores of individual grids, we compute the similarity for each grid separately. As a result, we not only know how many grids are similar, but also can obtain the list of similar grids. By the similarity score for each grid, we construct a similarity map for the scanning window. We name this map as a DOT similarity map.

Since the DOT similarity map is a binary image, the situation of DOT is quite different from HOG-LBP. Mean-Shift does not work with our DOT similarity map. In DOT representation, the energy function only indicates whether two grids are similar rather than how similar they are. If there exist some noises, two similar grids may not be matched. On the other hand, some different grids may be fortunately matched. In these situations, 0 or 1 will be returned and no intermediate values between 0 and 1 will be returned. If they are shown on the similarity map, we could seed some small holes. To fill these small holes, we apply the median filter on the DOT similarity map to remove these noisy grids. Based on the intensive experiments, the matched grids are usually dispersed in the false windows while they are usually concentrated in the true windows. From this observation, we can employ a very fast connected component detector to find the largest connected component in this map, which manifests the largest visible part of the query target. If the size of connected components is above an empirical threshold, we determine that the query target exists in this window.

3 Implementation Details

We summarize the details of the proposed occlusion handling approach in Fig. 1. To reduce the processing time, we only construct the DOT similarity maps for the windows of which the overall scores are larger than $T_w$. Therefore, lots of windows that are apparently not similar to the query target will be disposed directly. In our experimental settings, $T_w$ is usually smaller than the one used in the original DOT implementation in order to avoid missing some true regions with the small global scores. Then, we retrieve and refine the similarity map step by step according to the idea described in the previous subsection. In step E, we propose two criteria to measure whether a connected component should be a good detection. One is the area ratio of this component, which computes the ratio between the area of the component and the total area of the window. We denote it as the area score. The other is the similarity per grid inside this component, which calculates the ratio between the sum of similarities inside the component and the area of the component. We denote it as the valid score. Both of them are dependent on the thresholds $T_a$ and $T_v$. $T_a$ reflects how much occlusion the system can handle. Moreover, $T_v$ reflects how much noise the system can handle. If the grid is larger than $T_a$ and $T_v$, respectively, we determine that there may be a good detection in this window. This window is considered to be a candidate window. Finally, the decision is made by sorting the scores of candidate windows. We have evaluated the different scores for the sorting step. In the proposed approach, we employ the valid score while the global score is use in the original DOT implementation.
4 Experimental Evaluation

In our experiments, we use the PROST dataset [7]. Distance scores, PASCAL scores and detection rates are utilized to compare different approaches. Distance score calculates the mean distance of the tracking rectangle to annotated ground truth. The PASCAL score is defined as follows:

\[
score = \frac{\text{area}(\text{ROI}_D \cap \text{ROI}_{GT})}{\text{area}(\text{ROI}_D \cup \text{ROI}_{GT})}
\]  

(1)

The PASCAL score measures the overlap of the detected bounding box \( \text{ROI}_D \) and the ground truth bounding box \( \text{ROI}_{GT} \). Detection rate is defined as the ratio of frames in which the target is detected to the total number of frames in the sequence. Lower distance scores, higher PASCAL scores or higher detection rates indicate better performance. According to a lot of experiments, the parameters mentioned in Section 4 are listed as follows:

\[
\begin{align*}
T_w &= 0.6 \\
T_a &= 0.3 \\
T_v &= 0.8
\end{align*}
\]  

(2)

For the median filter, we set the radius to be 3. Due to the limitation of the length, we cannot show the experiments on achieving the optimal parameters.

First of all, we study the effectiveness of the proposed approach by comparing the object matching performance. The PASCAL score and distance score of each frame in sequences are plotted in Fig. 2. The average score of each sequence is shown on the right-top corner. The detection rates are shown in Fig. 3. Liquor and Tiger1 contain more occlusion cases than the Box sequence, and the improvements on both scores and detection rates are more significant. Box contained less occlusion cases and the improvement is marginal. Although the detection rate for Lemming is not improved with occlusion handling, we could find the locations of the target more exactly. Therefore, we get higher distance scores and PASCAL scores. Although there are few occlusions in Board and Dollar, the performance is still almost the same as the original DOT implementation. Additionally, some occluded regions that are successfully detected, as shown in Fig. 4.
speed of DOT is greatly dependent on the value of $T_w$. Significant speedup could be achieved by choosing a proper value for it. The experimental results demonstrate that DOT can achieve realtime performance with the proposed occlusion handling approach.

5 Conclusion

In this paper, we have presented a novel method to match objects with high robustness. We tackled the noise fragility and partial occlusions problems by calculating and analyzing the similarity maps to obtain the connected visible parts of the query object in sliding windows. The encouraging experimental results demonstrated that our method performs better than the original approach, especially on the occluded cases.

References