The High Repeatability of Salient Regions

Simone Frintrop

The High Repeatability of Salient Regions

Simone Frintrop

Institute of Computer Science III, University of Bonn, Germany

Abstract. In this paper, we show that salient regions, detected by a biologically motivated attention system, have a considerably higher repeatability than features detected by standard detectors (Harris-Laplacians and DoG features). If a scene contains a salient region, which differs in one of several possible features from the rest of the scene, the repeatability is close to 100%. In settings in which no especially salient region is present, the repeatability of attention regions is similar to DoG features and considerably higher than Harris-Laplacians.

1 Introduction

Feature detection is an important task for many applications in computer and robot vision. Usually, the detected features are used to establish correspondences between images. This is important in classical computer vision tasks like object recognition [1], wide baseline matching [2], and 3D reconstruction [3], but also in tasks for cognitive agents like robot localization [4], and simultaneous localization and mapping (SLAM) [5, 6].

Many good and stable feature detectors have been developed during the last decades (see [7] for a comparison). An important property of feature detectors is their *repeatability*: the same feature should be detected if a scene is visited from a different viewpoint or if the illumination changes. Usually, detectors compensate for a lack of repeatability by determining a large amount of features, which naturally increases the repeatability. However, in many applications a few stable features are sufficient and even preferable, since detecting, storing and comparing hundreds of features per frame is costly. Especially cognitive agents which have to share resources between different modules and tasks would profit from such a complexity reduction.

We show here that the human ability to focus on a few, salient regions in a scene can be exploited to achieve a sparse collection of features with very high repeatability. Biologically motivated attention systems mimic the mechanisms of the human visual system to determine the saliency of regions based on strong contrasts and uniqueness in several feature dimensions. Many attention systems have been developed during the last years [8–10] and have been evaluated with respect to their ability to simulate human eye movements [11]. They have been also used as feature detector for computer vision and robotics [6], but to our knowledge, they have not yet been compared to standard detectors. An exception is an entropy-based saliency detector [12] which was compared with other

detectors in [7].[1] It does however not consider the combination of different feature channels. In this paper, we show that regions with a high saliency have a considerably higher repeatability than features detected with standard detectors.

2 Biologically motivated salient regions

Biologically motivated attention systems base on the ability of humans to detect outliers in a scene. A black sheep in a white herd or a red ball on green grass attract our attention [13, 14]. Here, we use the attention system VOCUS [10]. It is in many aspects similar to the well-known system of Itti et al. [9] (a description of the differences can be found in [10]). One advantage of VOCUS is that it is real-time capable (50 ms for a 400×300 pixel image, on a 2.8 GHz PC [15]). Additionally, VOCUS has a top-down part to search for targets [10], but this is not considered here.

VOCUS creates a saliency map by computing image contrasts and uniqueness of a feature. The feature computations for the features intensity, orientation, and color are performed on 3 different scales with image pyramids. Two intensity feature maps, for on-off and off-on contrasts, are computed by *center-surround mechanisms*. Similarly, 4 orientation maps ($0\,^\circ, 45\,^\circ, 90\,^\circ, 135\,^\circ$) and 4 color maps (green, blue, red, yellow) are computed (cf. [10]).

Before the features are fused, they are weighted according to their *uniqueness*: a feature which occurs seldomly in a scene is assigned a higher saliency than a frequently occurring feature. This is the mechanism which enables humans to instantly detect outliers in a scene. The uniqueness $\mathcal{W}$ of map X is computed as $\mathcal{W}(X) = X/\sqrt{m}$, where m is the number of local maxima that exceed a threshold and '/' is here the point-wise division of an image with a scalar. The maps are summed up to 3 conspicuity maps I (intensity), O (orientation) and C (color) and combined to the *saliency map*:

$$S = \mathcal{W}(I) + \mathcal{W}(O) + \mathcal{W}(C) \tag{1}$$

The local maxima in S are sorted by saliency value (brightness in S), resulting in a ordered list of salient regions, the *VOCUS-ROIs*, $V_i = (V_1, ..., V_k)$. Note that V_j has a higher saliency than V_{j+i} ($i \in \{1, .., (k-j)\}$). The value of k is usually determined automatically, by considering all V_i that have a saliency value of at least $p\%$ (e.g. $p = 50$) of V_1. In the following experiments, we determine a fixed k to enable a direct comparison depending on the number of features.

3 Experiments

In our experiments, we show the high repeatability of salient regions and compare it with the repeatability of the standard detectors Harris-Laplace corners [16]

[1] The low score in [7] may be explained by the comparison method: a large amount of features (200-1600) per frame was considered, and the advantage of a saliency detector naturally shows off rather for a sparse feature representation.

		repeatability of strongest feature [%]		
data set	# frames	VOCUS-ROI	Harris-Laplacians	DoG features
1	400	100	8	21
2	305	95	32	53
3	259	97	44	59
4	291	70	47	78
5	209	82	12	74

Table 1. The repeatability of the strongest feature of ROIs from the attention system VOCUS, Harris-Laplace, and DoG features on the data sets illustrated in Fig. 1.

and Difference-of-Gaussians (DoG) blobs, i.e. extrema in DoG scale space [1].[2] To make the detectors comparable with the VOCUS-ROIs, we reduced the number of points by sorting them according to their response value and considering the same amount of features for each detector. We compared whether this response can be used to obtain a similar result as with salient regions.

As performance measure, we used the *repeatability*, similar to [7].[3] It is defined as the percentage of features (regions/points) in an image sequence which are detected in subsequent frames. All experiments were carried out on images of size 320×240.

3.1 Repeatability of strongest feature

In the first experiment, we evaluated the repeatability of the strongest/most salient feature. The experiments were performed on 5 image sequences of 200-400 frames (cf. Fig. 1). In all cases, strong viewpoint changes occur.

The 1st data set shows an artifically created scenery with many white and one green object with the same shape. The artificial set up enabled us to create a typical *pop out* case in which one object is considerably more salient than the others. Data set 2 and 3 show natural scenes in an office environment which contain objects which were especially designed to be salient for humans: a red circle containing a warning remark and a green exit sign. The last 2 data sets show scenes in which no especially salient object occurs, many regions are equally salient and are used for comparison.

Table 1 shows the repeatability of the most salient VOCUS-ROI V_1 and the strongest Harris-Laplace and DoG feature. For scenes containing especially salient objects (data sets 1–3), the repeatability of VOCUS-ROIs is, as to be expected, very high; it varies between 95 and 100%. The repeatability of Harris-Laplace and DoG features is considerably lower.

[2] We used the publically available PYRA real-time vision library for both Harris-Laplace and DoG features (http://www.csc.kth.se/∼celle/). Both detectors are, as usual, restricted to gray-scale images.

[3] In [7], also the number of correspondences were determined, but since we are interested in a low number, this performance measure is not useful here.

Fig. 1. The 5 data sets for the experiments in Tab. 1. 1st col: saliency map for 1st frame. 2-4th col: some frames with the most salient VOCUS-ROI V_1 (red ellipse). Complete sequences on http://www.informatik.uni-bonn.de/~frintrop/research/saliency.html).

Naturally, the advantage of attention regions especially shows off if there are salient objects. This is for example useful for a cognitive agent which is able to actively search for salient regions in its environment. Anyway, it is interesting to investigate what happens if there are no particularly salient objects but a normal, cluttered scene. We expected the VOCUS-ROIs to have about the same repeatability as the other detectors here. The last two rows of Table 1 show that the repeatability of VOCUS-ROIs and DoG points is indeed similar. Harris-Laplace corners have generally a much lower value. Note however that Harris-Laplacians have a very high accuracy which is important for some applications.

3.2 Repeatability of the k strongest features

In the second experiment, we investigate how the repeatability changes if not only one, but the k most salient features are considered. Naturally, the repeata-

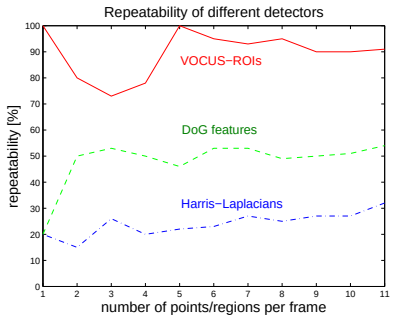

Fig. 2. Comparison of the repeatability of VOCUS-ROIs, Harris-Laplace, and DoG features on 10 frames of a sequence with a visually salient object (data set 1, Fig. 1).

bility goes up when considering more features.[4] However, since our concern here is to select few features, we consider small k.

We investigated the repeatability on the first 10 frames of data set 1 for different numbers of k. We investigated values up to $k = 11$, since this is the average value of VOCUS-ROIs detected in natural environments if k is chosen automatically (cf. sec. 2). The results are shown in Fig. 2. The highest repeatability is, as expected, obtained for $k = 1$, V_1 is on the same region in each frame. For $k \in \{2, 3, 4\}$, the VOCUS-ROI repeatability goes slightly down. This is easily explained: in the example, there is one especially salient object, the other objects are identical and have very similar saliency. Therefore, $V_2 - V_4$ tend to switch regions from frame to frame. For $k > 4$, the repeatability goes up again.

The repeatability of DoG and Harris-Laplace features is considerably lower, $20 - 30\%$ for Harris-Laplace and around 50% for DoG (except for $k = 1$). These first results illustrate the advantage of salient regions. It will however be necessary to repeat this experiment in future work for the other data sets. It is to be expected that the difference between curves is lower for natural scenes, especially for scenes without particularly salient objects. This goes conform with our claim: salient regions are highly repeatable, for scenes without salient regions the performance is comparable to standard approaches.

4 Conclusion

In this paper, we have shown the advantages of salient regions: they have a considerably higher repeatability than other image regions. Biologically motivated attention systems are able to capture this saliency from different feature

[4] Repeatability rates of about 60% have been reported for Harris-Laplacians in [7], with $k > 1000$.

dimensions. If a region pops out from the rest of the scene in at least one of 10 dimensions, its saliency is high. In applications in which a few correspondences between images are sufficient, this provides a significant advantage. One example is the loop closing in visual SLAM scenarios in which a previously seen location has to be redetected [6]. It would also be possible to actively search for expected, salient features by controlling the camera.

In a comparison with standard detectors, we have shown that the salient regions have a considerably higher repeatability. Naturally, this works best if the scene contains especially salient objects. In normal scenes without especially salient parts, the attention regions still show equal or better performance than the other detectors. The presented experiments are first results to illustrate the advantages of saliency. More detailed experiments are planned for the future to investigate how representative the values are in different settings.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int'l J. of Computer Vision (IJCV) **60** (2004)
2. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of BMVC. (2002)
3. Pollefeys, M., Gool, L.V., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a hand-held camera. Int'l J. of Computer Vision (2004)
4. Se, S., Lowe, D., Little, J.: Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. Int'l J. of Robotics Research **21** (2002)
5. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-Time Single Camera SLAM. Trans. on Pattern Analysis and Machine Intelligence **29** (2007)
6. Frintrop, S., Jensfelt, P.: Attentional landmarks and active gaze control for visual SLAM. IEEE Transactions on Robotics ((accepted))
7. Mikolajczyk, K., Schmid, C.: A comparison of affine region detectors. International Journal of Computer Vision (2006)
8. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial Intelligence **78** (1995) 507–545
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Trans. on Pattern Analysis and Machine Intelligence **20** (1998)
10. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-directed Search. PhD thesis (2005) Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer, 2006.
11. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research **42** (2002) 107–123
12. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Proc. of European Conf. of Computer Vision (ECCV). (2004)
13. Treisman, A.M., Gelade, G.: A feature integration theory of attention. Cognitive Psychology **12** (1980) 97–136
14. Wolfe, J.M.: Visual search. In Pashler, H., ed.: Attention. Hove, U.K.: Psychology Press (1998) 13–74
15. Frintrop, S., Klodt, M., Rome, E.: A real-time visual attention system using integral images. In: Proc. of Int'l Conf. on Computer Vision Systems. (2007)
16. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proc. of European Conf. of Computer Vision (ECCV). (2002)