Neurocomputing 120 (2013) 346–354

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An iterative combination scheme for multimodal visual feature detection

Gutemberg Guerra-Filho

Department of Computer Science and Engineering, University of Texas at Arlington, 500 UTA Boulevard, 76019-0015 Arlington, TX, USA

ARTICLE INFO

Article history:
Received 25 December 2011
Received in revised form
27 June 2012
Accepted 30 August 2012
Available online 27 March 2013

Keywords:
Feature detection
Multimodal detection

ABSTRACT

We address the problem of multimodal visual feature detection where several individual heterogeneous measures (*i.e.*, feature detectors) are merged into a single saliency value. We survey a number of techniques for the normalization and integration steps used in existing combination methods. A new approach, the iterative combination scheme, is proposed to iteratively learn a classifier that infers a non-linear model to combine different feature detectors. We evaluate and compare the combination strategies presented using an objective methodology, the repeatability criterion, and a dataset with real images of 21 cluttered scenes of 3D objects. Initially, our evaluation tested the performance of individual feature detectors. Considering the overall performance for all 7 scenes in the testing dataset, the Difference of Gaussian detector achieved the best repeatability rate, 54.41%. In our evaluation, we tested the performance of combining all possible sets of feature detectors. Among all possible sets, the triplet composed of Laplacian of Gaussian, Hessian Matrix, and Gradient Magnitude achieved the best performance of 58.93% repeatability. We used this combination of detectors to initialize the iterative combination scheme which was able to improve the performance to 66.62%.

1. Introduction

Primates analyze complex scenes efficiently by selecting a subset of the sensory input for processing [1]. This subset, named the focus of attention, is a small bounded region of the entire visual field. The selection of the attended region and the navigation through the visual field are based on the spatial distribution of saliency in the perceived scene. According to the feature integration theory [2], salient features in the visual scene are obtained by decomposing the visual input into a set of feature maps, detecting saliency within each map as spatial locations where discontinuities occur, and integrating all feature maps into a single saliency map. The saliency map, possibly found in the posterior parietal cortex of primates [3], encodes local features that are clearly discernible in the visual field by a scalar quantity.

Besides the empirical evidence in primates, the combination of multimodal features is supported by the hypothesis that a committee-based decision is usually better than a decision made by a single expert. Each different expert performs better on specific areas of knowledge. Therefore, the combination of expertise results in a decision system which considers a bigger domain of knowledge and performs better with regards to the quality of each decision.

The consistent combination of measures from different visual filters into a unique measure is an important step in the analysis and detection of features in low-level vision. The main challenge in combining heterogeneous measures is that they are not suitable for comparison. Since they are obtained by different methods, each measurement has a different range and a different distribution of its responses. Hence, the integration of diverse feature maps involves combining different modalities that are not comparable. In order to overcome this difficulty, the integration of heterogeneous measurements must consider the correlations between the different measures towards a unified inferred parameter.

A feature detector selects image locations presenting salient visual information. These low-level features are used in several tasks such as image indexing, shape reconstruction, stereo matching, object recognition, and others. Due to the relevance of low-level vision and to the numerous applications to artificial cognitive systems used in robotics, the literature on feature detectors is vast and a number of detectors have been proposed [4,5]. While these detectors are used individually with reasonable performance, little effort has been made to combine the power of different detectors into a single framework that achieves better results. In this paper, we address the issue of developing strategies to integrate heterogeneous saliency measurements into a single multimodal feature detector.

We propose a new approach to the multimodal feature detection problem, named iterative combination scheme (ICS). Our approach iteratively infers a classifier that combines a set of heterogeneous measures into a single saliency value. At each iteration, multimodal features are detected using the previously learned classifier. The features that are repeated in two images of the same scene are found. These repeated features are used to construct the input and the output data for the training of the next classifier. The initialization of our iterative combination scheme uses a combination method that consists of a normalization step and an integration step. We survey a

E-mail address: guerra@cse.uta.edu

0925-2312/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.neucom.2012.08.065

number of different techniques to address these two steps and, consequently, one of several combination methods (*i.e.*, pairs of normalization and integration steps) may be used in the initialization of the ICS procedure. This means that the first multimodal features are found using the best performing of these combinations of normalization and integration methods. After that, the next iterations will detect features using the previously learned classifier.

We evaluate the performance of our approach according to the repeatability criterion. Given two images of the same scene under different viewing conditions, repeatability is the fraction of detected features that are repeated in both images. We also consider accuracy and the cardinality of the set of detected features into the evaluation of the repeatability measure. With regards to accuracy, repeated features are defined according to a particular accuracy ϵ that represents the maximum distance between the detected feature and the actual feature obtained from a ground truth disparity map. The cardinality of detected features is limited to a fixed percentage of the number of pixels in each image. This way, different detectors will result in a equal number of detected features for the same image.

While saliency selects a reduced amount of visual information (compactness), the relevance of the repeatability measure concerns the ultimate goal of corresponding features in different images which impacts several vision issues (reconstruction, tracking, recognition). Hence, repeatability is the quantitative measure used to evaluate the performance of point feature detectors with respect to a particular accuracy and compactness levels. Our major goal using repeatability is to provide a fair evaluation where all pixels present in both images are considered. Thus, we consider a dense evaluation not influenced by any biased selection of a sparse set of features.

The images used in our evaluation are in the 2006 Middlebury dataset [6]. The Middlebury dataset provides greater 3D shape variation than planar scenes in other datasets. The planar scenes in these datasets may be biased towards affine feature detectors specifically designed to take homographies into consideration. For this reason, we avoid an evaluation that considers only planar scenes. The Middlebury dataset contains images of 21 scenes cluttered with a diverse set of 3D objects under three different lighting conditions, three exposure levels, and three different resolutions. The ground-truth for disparity maps between pairs of stereo images is available for this dataset. Hence, since pixel correspondence between different images is given, the exact computation of the repeatability rate for a feature detector is feasible. Furthermore, using these disparity maps, we reconstruct the 3D scene and reproject it according to different general camera poses. These images of real scenes obtained from the 3D reconstruction are also used in the evaluation of our method. We compare the performance of individual feature detectors with the performance of our iterative combination scheme. The average performance of our multimodal feature detector in the test image dataset is 66.62% repeatability ratio while the best individual detector, the Difference of Gaussians, has a 54.41% repeatability ratio. Hence, ICS improves on the best individual detector by an additional 12.21% repeatability.

The main contributions of this paper are: (1) a novel method, the iterative combination scheme, for the combination of heterogeneous measures into a single multimodal feature detector that outperforms state-of-the-art detectors with regards to repeatability, (2) a survey with a number of different techniques to address the normalization and integration steps of a combination approach, (3) the quantitative evaluation and comparison of a set of well-known feature detectors (Gradient Magnitude, Harris Corner, Hessian Matrix, Difference of Gaussian, and Laplacian of Gaussian) using cluttered scenes of 3D objects with ground-truth available for the exact computation of repeatability rates, and (4) evaluation of several strategies for the combination of feature detectors into a single multimodal measure and its comparison to the individual detectors.

The remaining of this paper is organized as follows. In Section 2, we formalize the multimodal feature detection problem as the combination of heterogeneous measures and we discuss several techniques for the normalization and integration steps of a combination method. We present our novel iterative combination scheme in Section 3 by describing the general framework and the specific details on the classifiers considered in our implementation (Artificial Neural Networks). In Section 4, we review the previous work on the evaluation of feature detectors. The evaluation criteria used in our experiments is described in Section 5.1. In Section 5, we describe several experiments to assess and compare the performance of individual feature detectors and of the different strategies for the combination of detectors into a single multimodal measure of saliency. Our conclusions in Section 6 elaborate on our findings about combination of feature detectors and on possible future directions.

2. Combination of heterogeneous measures

Assume that we have a set of n feature detectors $\{F_1,F_2,\ldots,F_n\}$. Each detector F_i is a function $y_{i,p}=F_i(x_p)$, where $y_{i,p}\in\mathbb{R}$ is the saliency measure associated with detector F_i and $x_p\in\mathbb{R}^s$ is an input vector obtained from a rectangular window of size s centered at pixel p in the image. The detector F_i processes an input vector x_p to obtain the level of support $y_{i,p}$ to the hypothesis that the corresponding pixel is a salient feature. The combined output of all detectors is a n-dimensional vector $Y(x_p)=[F_1(x_p),F_2(x_p),\ldots,F_n(x_p)]$. The combination of detectors is formally defined as integrating the vector $Y(x_p)$ into a single measure $y(x_p)\in\mathbb{R}$.

A *naïve combination* of the heterogeneous measures in vector $Y(x_p)$ into a single measure $y(x_p)$ involves two steps: normalization and integration. The normalization and integration framework was independently used by many previous work that addressed the merging of heterogeneous measures into a single value [7]. The normalization step transforms all feature maps into a single range where comparisons and operations to integrate the different measures are suitable. After the normalization step, all feature maps are theoretically transformed to the same commensurate range. The integration step combines the normalized values of all feature detectors into a single value that represents a multimodal measurement of saliency in the image.

The normalization and integration steps may be applied to a pyramid of images in scale-space generated from the original image. Each feature map is computed at each scale by a procedure akin to visual receptive fields. After normalization and integration at different scales, the feature maps may be combined through across-scale addition. Fig. 1 illustrates the combination method for a single image scale.

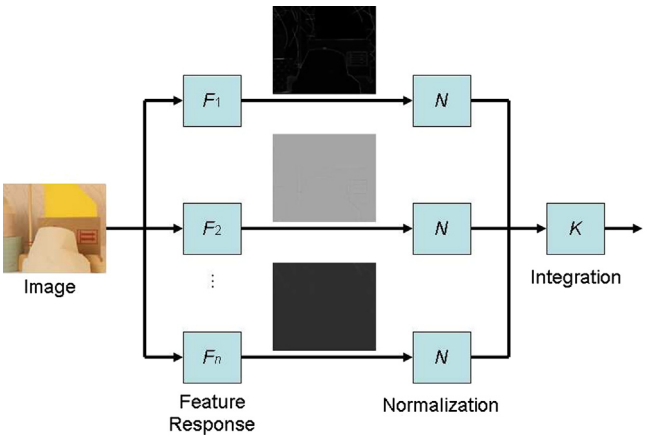

Fig. 1. The combination method with normalization and integration steps.