ABSTRACT OF THE DISSERTATION

Semantic Image Representation for Visual Recognition

by

Nikhil Rasiwasia

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2011

Professor Nuno Vasconcelos, Chair

A novel image representation, termed semantic image representation, that incorporates contextual information is proposed. In this framework, images are represented by their posterior probabilities with respect to a set of appearance based concept models, built upon the bag-of-features representation. Thus while appearance features are intensity, texture, edge orientations, frequency bases, etc. those of the semantic representation are concept probabilities. Semantic image representation induces a mapping from the space of appearance features to a *semantic space*, where each axis represents a semantic concept. Each concept probability is referred to as a *semantic feature* and the semantic feature vector as the *semantic multinomial* (SMN) distribution. Next, we present design of three different

visual recognition tasks viz. image retrieval, scene classification and cross-modal multimedia retrieval, based on the semantic image representation. First, a novel framework for content based image retrieval, referred to as *query by semantic example* (QBSE) is proposed, which extends the query-by-example paradigm to the semantic space. Current content based image retrieval solutions rely on strict visual similarity, which in most cases, is weakly correlated with the measures of similarity adopted by humans for image comparison. By using the semantic image representation, the retrieval operation is performed at a much higher level of abstraction, which results in retrieval systems that are more accurate than previously possible. QBSE also allows a direct comparison of visual and semantic representations under a common query paradigm, which enables an explicit test of the value of semantic representations for image retrieval. Second, we propose a framework for scene classification based on the semantic image representation. As in previous approaches, we introduce a low dimensional intermediate space, which in the proposed framework is served by the semantic space. However, instead of learning the intermediate "themes" in an unsupervised manner, they are learned with weak supervision, from casual image annotations. When annotations are not available, they are replaced by the scene category labels. A study of the effect of dimensionality on the classification performance is also presented, indicating that the dimensionality of the "theme" space grows sub-linearly with the number of scene categories. Third, the problem of cross-modal retrieval from multimedia repositories is considered. This problem addresses the design of retrieval systems that support queries *across* content modalities, *e.g.*, using text to search for images. A mathematical formulation is proposed, where the design of cross-modal retrieval systems is equated to that of designing isomorphic feature spaces for different content modalities. Three new solutions to the cross-modal retrieval problem are proposed: correlation matching (CM), which models cross-modal correlations between different modalities, semantic matching (SM), which relies on the semantic representation, where different modalities are represented on a common semantic space, and semantic correlation matching (SCM), which combines both. An implementation of the above systems under the minimum probability

of error framework is presented and compared to various existing algorithms in respective visual recognition tasks, on benchmark datasets. It is shown that the proposed semantic image representation is able to achieve superior results. Finally, we discuss the issue of *contextual noise* in semantic representations, due to the inherent ambiguity of the bag-of-features representation. To address this, we propose a novel two-layer framework to context modeling, based on the probability of co-occurrence of objects and scenes. The first layer represents the image in a semantic space, and the second layer introduces distributions of each concept in the semantic space. This facilitates robust inference in the presence of contextual noise. A thorough and systematic experimental evaluation of the proposed context modeling is presented. It is shown that it captures the contextual "gist" of natural images. The effectiveness of the proposed approach to context modeling is further demonstrated through a comparison to existing approaches on scene classification and image retrieval, on benchmark datasets. In all cases, the proposed approach achieves state of the art visual recognition performance.