Small codes for large systems: information-theoretic privacy preserving content identification

Emmanuel Abbe, Princeton University
Svyatoslav Voloshynovskiy, University of Geneva

Abstract
In recent years, bag-of-word (BOW) based recognition, classification and retrieval systems have become the state-of-the-art in many applications ranging from multimedia management to security. In addition, in many applications facing strict memory-complexity restrictions, like in the on-line mobile phone visual or audio search systems, the BOW systems with carefully designed descriptors probably remain the only suitable technology in comparison to emerging yet complex deep learning frameworks. The core idea behind the BOW systems consists in a representation of each image, described by a set of short and local descriptors, by a fixed dimensional feature vector that is invariant to geometrical de-synchronization.

However, besides of its remarkable experimental performance, the theoretical analysis of BOW based systems especially when applied to visual content identification remains largely unexplored. These systems are often considered as black boxes where the performance is estimated based on a public database, for some type of descriptors (e.g., SIFT, SURF, CHOG, ORB, BRIEF, etc.) with little theoretical insight. In addition, it is not completely clear which descriptors in which particular situation contribute to successful identification. Moreover, if the descriptors are changed, the systems performance should be assessed again via time-consuming simulations.

Further more, in security and privacy applications, it is not obvious which factors determine the security of these systems and which elements of the BOW systems should be properly protected. Finally, the optimality of different encoding/assignment and pooling methods is based on empirical evidence rather than on strictly proven theoretical results.

Brief overview of state-of-the-art BOW-based content identification
Currently, most BOW systems are used for CBIR, object recognition and copy detection. We will consider *content identification* where M items are enrolled and given a probe, the system should determine the corresponding item or issue a rejection. When it is not possible to return a single index item, the system should retrieve a list of indices whilst ensuring that the true item index is on the list. The CBIR counterpart of content identification retrieves a list of indices of items similar to the probe.

The performance of BOW-systems is generally evaluated by simulation, and existing theoretical works [1-4] mostly consider content identification based on content fingerprinting where a sufficiently long fingerprint is deduced to represent the content. In most theoretical works, perfect synchronization between the enrolled

fingerprint and the probe fingerprint is assumed with one notable exception [1] where fingerprint de-synchronization was modeled by a random shift parameter. However, in practice it is not feasible to design one single fingerprint or descriptor that would be invariant to all types of distortions, hence multiple local descriptors or small codes per image are used. Despite popularity, SIFT descriptors are characterized by the high computational complexity and relatively long length. In the recent years, a number of short binary descriptors have been proposed (BRIEF, ORB, etc). However, in this case the length of the deployed descriptors does not satisfy the asymptotic assumptions considered in the theoretical works [1-4] which makes the analysis of practical BOW-systems intractable.

Scope and goal of proposed project

The proposed project will focus on the information-theoretic analysis of BOW based content identification systems. To our knowledge there is little work on the theoretical analysis of BOW-systems' performance besides [5] and none on BOW based content identification. In addition, the privacy-preserving aspects of BOW based systems remain an emerging and little studied problem. In [6], it was shown that privacy-preserving computations can be made efficient for specific functionals related to statistical estimation, and in [7] the construction of new polar coding schemes are shown to achieve the secrecy capacity with low complexity. The project will leverage these various techniques in the context of BOW based systems. The recent work on linear classification [8] will also be studied in this context.

The project applicants pose all necessary skills and have proven records of achievements to address the above problems. Therefore, the goal of this problem is to provide a simple and tractable model allowing to analyze, optimize and guide the design of BOW systems. We will consider the case of non-compressed features to reveal the theoretical limits of BOW based identification systems, analyze the impact of descriptor compression and encoding/assignment as well as discovering the impact of geometrical consistency between the descriptors on overall system performance. Such a formulation was not considered in earlier studies.

Organization of collaboration

Upon approval the project applicants will develop a plan of exchange visits and seminars. In view of limited budget, skype meetings will also be used for the scientific discussions, and we will try to use all opportunities to meet during international conferences (in particular, the bi-annual Conference on Information Sciences and Systems at Princeton University in March). On-site visits will be planned for the kick-off meeting and presentation of results in the host institutions. The invited seminars will be given by the project applicants to the Faculty members, researchers and students thus disseminating the results of our collaboration.

We intend to engage our PhD students in active exchange of ideas and will foster joint publications. Depending on the budget, we also plan several exchange visits of PhD students. We will use any opportunity to present our results to industrial partners to attract the industrial funding and grants. Finally, we will promote this collaboration while applying for the National Science Foundation projects, both in Switzerland and in the USA.

References

1. P. Moulin, “Statistical modeling and analysis of content identification,” in *Information Theory and Applications Workshop (ITA), 2010*. IEEE, 2010, pp. 1–5.
2. A. L Varna and M. Wu, “Modeling and analysis of correlated binary fingerprints for content identification,” *Information Forensics and Security, IEEE Trans- actions on*, vol. 6, no. 3, pp. 1146–1159, 2011.
3. F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, “Performance analysis of content- based identification using constrained list-based decoding,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1652–1667, 2012.
4. S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak, “Information- theoretical analysis of private content identification,” in *IEEE Information Theory Workshop, ITW2010*, Dublin, Ireland, Aug.30-Sep.3 2010.
5. Yin Zhang, Rong Jin, and Zhi-Hua Zhou, “Understand- ing bag-of-words model: a statistical framework,” *Inter- national Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.
6. E. Abbe, A. Khandani, A. W. Lo, “Privacy-preserving methods in systemic risk", Proc. and Papers of the American Economical Review (AER), 2012. New York Times article: http://bits.blogs.nytimes.com/2013/09/09/a-data-weapon-to-avoid-the-next-financial-crisis/
7. E. Abbe, “Extracting randomness and dependencies via polarization, Slepian-Wolf coding and secrecy” presented at ITA, Feb 2011, to appear in the IEEE Information Theory Transactions.
8. E. Abbe, N. Alon, A. Bandeira, “Linear coding, classification and the space-avoiding-set problem”, preprint, 2013.