VPSLAM: VISUAL PROBABILITY MASS FUNCTION SLAM

Clebson J. M. de Oliveira, Raquel F. Vassallo, Evandro O. T. Salles*

*VIROS/LABCISNE, Engenharia Elétrica, Universidade Federal do Espírito Santo (UFES)

Email: clebson.oliveira@aluno.ufes.br, raquel@ele.ufes.br, evandro.salles@ufes.br

Abstract— We present a novel Visual Simultaneous Localization and Mapping (VSLAM) algorithm, the VPSLAM algorithm (Visual Probability Mass Function SLAM), and tested it using some image sets from the state of the art in SLAM methods. The obtained results show that the proposed VPSLAM algorithm can perform place recognition in outdoor environments and is able to detect loop closures, even in low resolution imagery. Also, we can conclude that the characterization model C-IGL can be applied in appearance-based SLAM algorithms over dynamic environments.

Keywords— Visual SLAM, Probabilistic Model, Outdoor Environment.

1 Introduction

The global localization of a mobile robot can be defined as an instance of the general localization problem, which is a fundamental perception problem in Robotics. In fact, almost all tasks in Robotics need the location of the object of interest, no matter if it is an object transported by the robot or the robot itself.

Currently, a common approach to solve the outdoor visual navigation problem is to use Visual Simultaneous Localization and Mapping (VSLAM) algorithms.

In those algorithms, each location in the robot's operating environment is defined by descriptors extracted from it. Then, the localization is performed by matching the current image's descriptors and the descriptors assigned to each place in the environment.

Johannsson and colleagues presented a system for temporally scalable visual SLAM (Johannsson et al., 2013) where new measurements are used to continually improve the map. That approach achieves efficiency by avoiding adding redundant frames and not using marginalization to reduce the graph. The presented results demonstrate long-term mapping in large-scale. However, the experiments were restricted to indoor mapping.

Zhou et al. (2015) presented the StructSLAM, a novel 6-degree-of-freedom visual SLAM method based on the structural regularity of man-made building environments. The base idea is that the building structure lines are used as features for localization and mapping. The results show that StructSLAM performs remarkably better than the existing methods in terms of position error and orientation error. However, it is based on man-made information.

In this work, we developed a VSLAM algorithm based on the appearance-based characterization model C-IGL (de Oliveira et al., 2015) and performed experiments in outdoor environments. So, this paper presents a new VSLAM algorithm which can be applied to real non-supervised learning place recognition problems.

2 Background

Considering the appearance-based approach for localization, one of the most significant contributions is the algorithm *Fast Appearance-Based Mapping* (FAB-MAP) (Cummins and Newman, 2008) and its further improvement FAB-MAP 2.0 (Cummins and Newman, 2010). Such work is a new formulation of Simultaneous Localization and Mapping (SLAM), and it was very successful because of the tests performed over large distances.

Despite the performed experiments had been successful, the algorithm had to be trained offline to learn a global features vocabulary.

A significant drawback of place recognition models based on FAB-MAP is the determination of a global features vocabulary, which can be an obstacle to implement real-time application. This limitation occurs because it is necessary to train the vocabulary along the entire environment. So, alternative models are desired.

Milford (2013) discusses the current trend of using a huge amount of visual information in appearance-based applications. Milford performs studies to define an upper-bound limit of small and poor information used in SLAM methods.

The experiments in Milford (2013) overcame the FAB-MAP performance using the SeqSLAM method (Milford and Wyeth, 2012) - a visual SLAM algorithm which basically compares sequence of images, instead of single images, to define the correspondences.

Milford also discussed the influence of visual information parameters in SLAM methods like field-of-view (FOV), resolution and blurring. Considering the challenging data sets tested in (Milford, 2013) we decided to test our model in some of them also.

3 The Probability Mass Function SLAM: VPSLAM

In this section we present a new SLAM algorithm based on the proposed model C-IGL (de Oliveira et al., 2015). Our goal is to explain how the C-IGL model can be applied in a real-time SLAM algorithm.

For that, we consider each new sample S linked to an instant of time t and calculate $p(l_i|S_t)$ using the C-IGL model. Thus, the localization method along time is,

$$X(S_t) = \underset{i}{\operatorname{argmax}}(p(l_i|S_t)), \forall l_i \in L_t, \qquad (1)$$

where L_t is the set of places mapped until the instant t. To use (1) in a SLAM algorithm it is necessary to identify possible loop closure - a visit to previously mapped places. So, we used the thresholding technique in Algorithm 1.

Input: S_t, L_t, K
Output: L_{t+1}
begin
  $l_{nn} = \underset{i}{\operatorname{argmax}}(p(l_i|S_t))$
  if $p(l_{nn}|S_t) <= K$ **then**
    /*Add sample S_t into the map*/
    $L_{t+1} = addPlace(S_t, L_t)$
  else
    /*The map remains the same*/
    $L_{t+1} = L_t$
  return L_{t+1}

Algorithm 1: VPSLAM Algorithm.

For each new sample S_t we find its nearest neighbor l_{nn} and try to add the place by comparing its similarity probability to a probability threshold K, as defined in Algorithm 1.

If S_t is not considered similar to any place in the previous map L_t, the new map of the environment L_{t+1} will present a new place.Then, the function $addPlace(S_t, L_t)$ adds a new place in L_t using the current sample S_t.

The *argmax* computation in the Algorithm do not search in the last *ir* places of the map. The *ir* places are probably more similar to the current place than a previous visit to the current place. In addition, it is improbable to find a previous visit to the current place in the last places added to the map.

It is important to point that the *ir* parameter has the same effect in the algorithm performance as the sequence length parameter in SeqSLAM (Milford and Wyeth, 2012).

The sequence length parameter in SeqSLAM provides a benefit to the algorithm performance, as discussed in Milford and Wyeth (2012), because the sequence matching is determined by the correspondent sequence center. So, it is not determined the best nearest neighbor (NN) but a good one.

4 Experimental Results

Considering the state of the art in place recognition one can cite the work developed by Milford and colleagues (Milford, 2013). This paper presents experiments over different image sets to test influences like resolution reduction, FOV, blurring, image sub-sampling, pixel bit depth and others.

We have tested the VPSLAM algorithm using some of the challenging image sets used in (Milford, 2013), *Rowrah*, *Mt Cootha* and *Pike Peaks*. We also used two homemade imagery sets, that we called *Around the block front* and *Around the block aside*. Each image set used in our experiments consists of two traverses in the same environment to test different image parameters as follows:

- *Rowrah*: low resolution imagery captured in a motor bike speedway;
- *Mt Cootha*: 630 ms exposure-time imagery, one traverse captured in the afternoon and one traverse in the the middle of the night;
- *Pike Peaks*: low quality imagery captured in a rally race;
- *Arround the block front*: high quality imagery along 900 meters of an urban block with the camera mounted on the car's panel and pointed at the street in front of the car.
- *Arround the block aside*: high quality imagery along 900 meters of an urban block with the camera mounted on the car's side and pointed at the street's margin.

Rowrah and *Pike Peaks* were sub-sampled in comparison to the frame rate in the video. This was necessary to obtain image amounts comparable to the one used in (Milford, 2013).

4.1 Evaluation Methodology

To test each image set we calculate $p(l_i|S_t)$ in the Algorithm 1 using the C-IGL model (de Oliveira et al., 2015). The parameter K was set to zero and then we built the precision-recall curve to evaluate the loop closure detection performance.

Setting the parameter K to zero is necessary to evaluate the algorithm independently to the parameter choice, once every new sample S_t is added to the map and the precision-recall curve is generated sweeping the threshold K over the range $[0, 1]$.

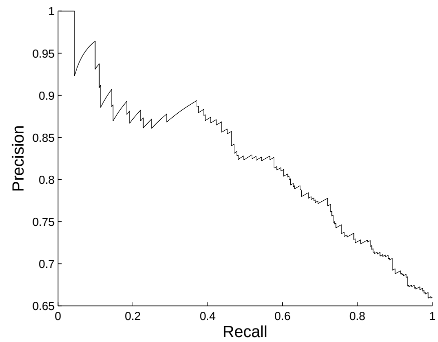
For building the precision-recall curves it is necessary to define the ground truth for each image set - to label (true or false) the nearest neighbor defined by VPSLAM for each image. Once the image sets from (Milford, 2013) do not present GPS data, Milford and colleges have defined the ground truth by analysing each sequence matching, i.e. NN definition, based on human visual perception.

Nevertheless, human visual perception is easily tricked by degradation on image parameters such as blurring, image capture at night (*Mt Cootha*), low resolution imagery (*Rowrah*) and low quality imagery (*Pike Peaks*).

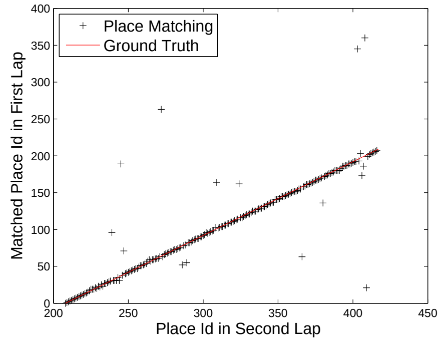
Thus, we have defined automatically the ground truth. Our procedure to automatically label the NN is based on the assumption that the camera moves approximately at the same average speed while comparing the second and the first traverse at the environment.

Then, the place l_i in the first traverse is the nearest neighbor of the place l_j in the second traverse when $i = j$. However, it is necessary to consider an uncertainty parameter $resolution$ to equalize the speed assumption.

So, to build the precision-recall curves we considered a correct NN definition when $abs(i - j) = resolution$. For places in the first traverse l_i we considered the NN when the matched place's identification (id) is greater than $i - ir - resolution$.



(a) $ir = 5$ - Precision Recall.



(b) $ir = 5$ - Place Matching.

Figure 1: Evaluation using low resolution/quality images (*Rowrah*). One image per place and $ir = 5$.

4.2 Results

To be able to compare VPSLAM with SeqSLAM over the same image sets we have based our analyses on the same metrics presented in (Milford, 2013). The precision-recall curves demonstrate the loop closure detection performance, i.e. the capability of the SLAM algorithm to recognize previously mapped places.

To identify the precision of the algorithm we present graphics showing the nearest neighbor of each place in the second traverse, i.e. the place matching.

Rowrah image set presented a good performance for $ir = 50$, but it is possible to see the improvement caused by the increasing in parameter ir in Figures 1 and 2.

The precision rate do not reach 100%, but for $ir = 50$, Figure 2a, it is greater than 90% and the recall rate reaches 100% at precision near 60%.

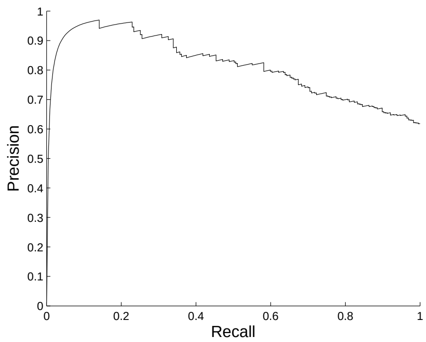
Figures 1 and 2 show that even for a low ir value equal 5 the place matching is precise. An example of this place matching definition can be observed in Figure 3.

The result in Figure 1b is similar to the one in Figure 4, presented by Milford and colleagues in (Milford, 2013), where the vertical axis shows the central frame number of the matching sub-route from the first traverse of the route.

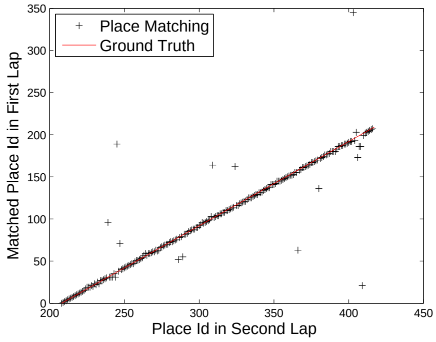
(a) Recalled image in first lap.

(b) Image in second lap matched to image 3a.

Figure 3: Image matching in *Rowrah* for $ir = 5$.



(a) $ir = 50$ - Precision Recall.



(b) $ir = 50$ - Place Matching.

Figure 2: Evaluation using low resolution/quality images (*Rowrah*). One image per place and $ir = 50$.

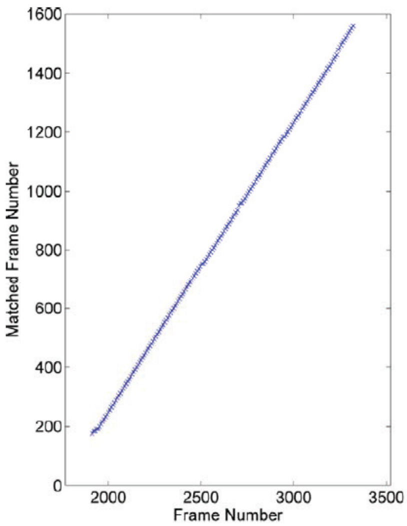
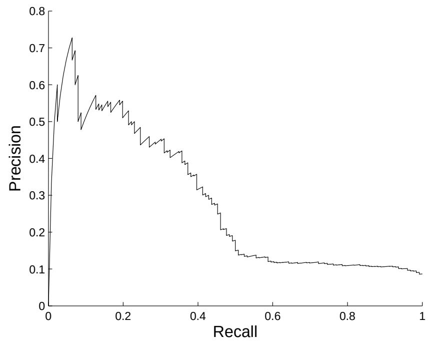


Figure 4: Sequence matching on *Rowrah* data set presented in Milford (2013).

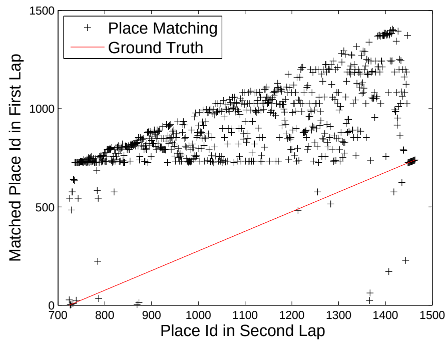
Mt Cootha image set is really challenging because of the existence of night images and also the long exposure time, 630 ms, used to capture the images.

In our experiments the best result of *Mt Cootha* is obtained for $ir = 10$. Figure 5a shows that 100% precision is never achieved in our experiments, but in the experiments performed in (Milford, 2013) the 100% precision is always achieved even for different exposure time, between 630 ms and 10,000 ms.

However, Figure 5b shows that VPSLAM can match images captured in extremely different illumination condition, because there is a few matchings equal to the ground truth.



(a) $ir = 10$ - Precision Recall.



(b) $ir = 10$ - Place Matching.

Figure 5: Evaluation using blurring and night images (*Mt Cootha*). One image per place and $ir = 10$.

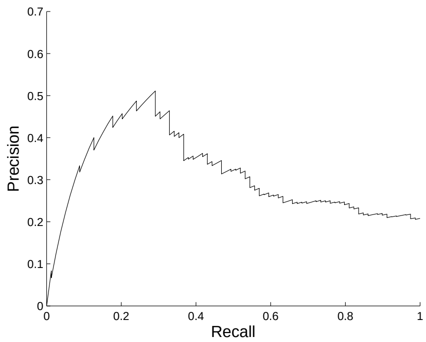
Thus, the results presented in Figure 5 show that the characterization model and consequently the algorithm VPSLAM presented in this paper, are not invariant to extreme illumination conditions, once just a few images were correctly matched.

SeqSLAM's results presented in (Milford, 2013) show that is possible to match images in different blurring conditions and between day and night imagery. So, we conclude that VPSLAM is not feasible for image comparisons between day and night because we never achieved 100% precision and the image matching is not correct for the majority of the tested places, Figure 5b.

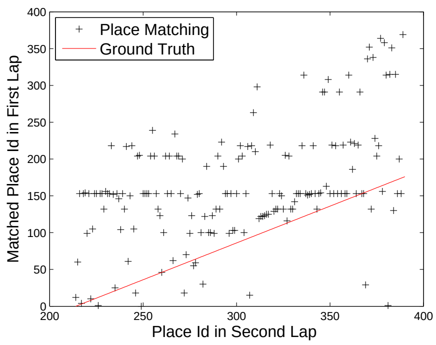
Pike Peaks presents images from a rally race captured at pixel resolution 960x720, in the first traverse, and 1280x720 in the second traverse. This condition can degrade the performance of VPSLAM once the place captured in different pixel resolutions can change the SURF descriptor of an interesting point.

That characteristic can be confirmed by Figure 6a, where 100% precision is never achieved. In Figure 6b almost all places in the second traverse are incorrectly matched to places in the first traverse.

Although VPSLAM is not robust to differences in pixel resolution, this is not a real robustness parameter to be available in the SLAM context, because in practice the camera will be the same along the entire path.



(a) $ir = 10$ - Precision Recall.

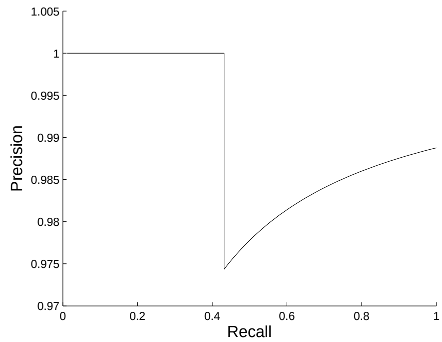


(b) $ir = 10$ - Place Matching.

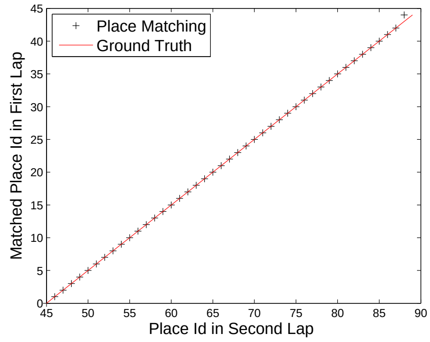
Figure 6: Evaluation using pose variant images (*Pike Peaks*). One image per place and $ir = 10$.

Around the block image sets, *front* and aside, were captured to test VPSLAM over a high quality image set. The idea is to observe the potential of the algorithm without forcing degrading capture conditions.

Figure 7a shows it is possible to achieve 100% recall rate considering recently mapped places, i.e. $ir = 0$, and Figure 7b demonstrates the VPSLAM's capability to precisely determine place matching.



(a) $ir = 0$ - Precision Recall.



(b) $ir = 0$ - Place Matching.

Figure 7: Evaluation using the camera pointed at the street (*Around the block front*). One image per place and $ir = 0$.

We also aim to test a hypothesis: the camera pointed at the street in front of the car can degrade the characterization power of the proposed model.

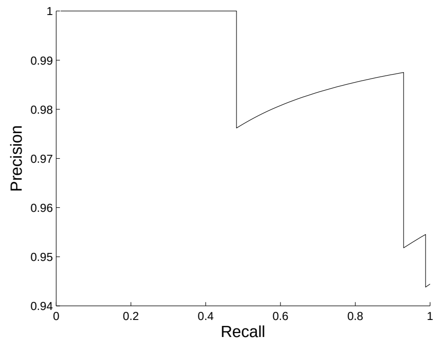
This hypothesis came up because our model is based on invariant SURF features. So, when we capture the same place changing scale is really probable to overlap the extracted features in different places, which is not desirable for place matching.

Therefore, we traveled the same path of *Around the block front* but pointing the camera at the street's margin, *Around the block aside*. Figures 8a, 9a show that the recall rate can rises when the parameter ir rises. For $ir = 0$, the maximum recall rate at 100% precision, 48.81%, is greater than the maximum recall rate at 100% precision when the camera is pointed at the street, 43.18%.

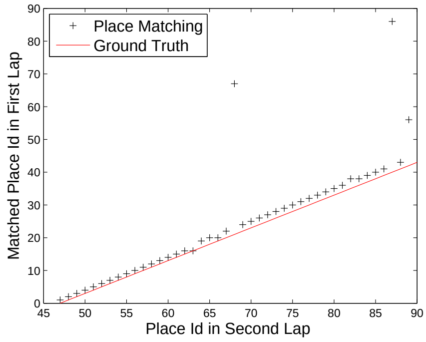
The results presented in Figure 9 show that VPSLAM presents better performance when the camera is pointed at the street's margin and the maximum recall rate at 100% precision can be comparable to the performance of SeqSLAM.

5 Conclusions

Considering the problem of navigating in outdoor environments we proposed a visual SLAM algorithm named VPSLAM, which learns places at

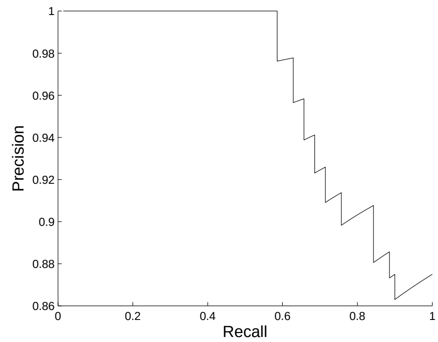


(a) $ir = 0$ - Precision Recall.

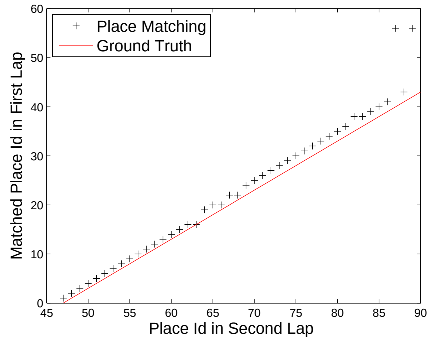


(b) $ir = 0$ - Place Matching.

Figure 8: Evaluation using the camera aside (*Around the block aside*). One image per place and $ir = 0$.



(a) $ir = 10$ - Precision Recall.



(b) $ir = 10$ - Place Matching.

Figure 9: Evaluation using the camera aside (*Around the block aside*). One image per place and $ir = 10$.

a fixed rate and can detect loop closures in the mapped environment. VPSLAM is an application of the C-IGL model which can be applied to non-supervised learning place recognition algorithms. So, our work can be used to solve real SLAM problems in a feasible approach.

The main contribution of the proposed VP-SLAM is not to use a global vocabulary as commonly used by SLAM algoritms. The removal of the pre-processing step to generate the global vocabulary, proposed here, can improve the computational efficiency obtained by using local feature vocabularies.

An important step to be done is to adapt the model to handle extreme illumination conditions like localization between day and night. This future work can improve the capability of the model to distinguish illumination noises and features in extreme intensity levels.

References

Cummins, M. and Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance, *The International Journal of Robotics Research* **27**(6): 647–665.

Cummins, M. and Newman, P. (2010). Appearance-only slam at large scale with fab-map 2.0, *The International Journal of Robotics Research* .

de Oliveira, C. J. M., Vassallo, R. F. and Salles, E. O. T. (2015). Pre-filtering effect on visual initial global localization, *Anais do XII Simposio Brasileiro de Automacao Inteligente (XSBAI)*, pp. 1188 – 1193.

Johannsson, H., Kaess, M., Fallon, M. and Leonard, J. J. (2013). Temporally scalable visual slam using a reduced pose graph, *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 54–61.

Milford, M. (2013). Vision-based place recognition: how low can you go?, *The International Journal of Robotics Research* **32**(7): 766–789.

Milford, M. and Wyeth, G. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights, *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 1643–1649.

Zhou, H., Zou, D., Pei, L., Ying, R., Liu, P. and Yu, W. (2015). Structslam: Visual slam with building structure lines, *IEEE Transactions on Vehicular Technology* **64**(4): 1364–1375.