A&A 405, 1169–1176 (2003)
DOI: 10.1051/0004-6361:20030564
© ESO 2003

Astronomy & Astrophysics

Increasing the reliability of ISOCAM[⋆] cross-identifications by use of a probability pattern

S. Derriere[1], S. Ott[2], and R. Gastaud[3]

[1] Observatoire Astronomique de Strasbourg, 11 rue de l'Université, 67000 Strasbourg, France
[2] ISO Data Centre, Science Operations and Data Systems Division, European Space Agency, Villafranca del Castillo, PO Box 50727, 28080 Madrid, Spain
e-mail: sott@iso.vilspa.esa.es
[3] CEA/DSM/DAPNIA Saclay, Service d'Astroph., L'Orme de Merisiers, Bat. 709, 91191 Gif-sur-Yvette, France
e-mail: rgastaud@cea.fr

Received 16 May 2002 / Accepted 13 March 2003

Abstract. The built-in play at the wheels of the ISOCAM instrument affects the accuracy of the astrometry of sources from the ISOCAM Parallel Catalogue. We present the results of different methods for cross-identifying ISOCAM sources with reference catalogues in the optical (USNO-A2.0) and in the near infrared (2MASS). For a sample of 7197 ISOCAM sources, the fraction having a real counterpart in the USNO and 2MASS catalogues was estimated to be 95% and 91% respectively. Using a nearest neighbor method for cross-identification, only 74% of the possible associations with USNO (79% with 2MASS) are retrieved. With the use of a so-called probability pattern to perform the associations, these values increase to 77% for USNO and 83% for 2MASS. In addition, the overall reliability of associations increases from 77 to 87% with USNO, and from 81 to 90% with 2MASS.

Key words. catalogs – astrometry – methods: data analysis – methods: statistical – infrared: general

1. Introduction

Onboard the Infrared Space Observatory (ISO) satellite (Kessler et al. 1996), the mid-infrared camera ISOCAM (Cesarsky et al. 1996) continued observing the sky while another instrument was prime. From these "parallel mode" observations (Siebenmorgen et al. 1996), covering 47 square degrees, a catalogue has been built (Ott et al. 2003), consisting mainly of observations in broad band filters centered around 6 μm, with a 6″ pixel field of view (PFOV), with fluxes down to 0.5 mJy.

The astrometric accuracy of the 6″ PFOV observations is mainly affected by three factors:

- a 2″ pointing accuracy for ISO;
- up to 1 PFOV distortion at the edges of the detector (this effect is already corrected in the catalogue);
- a shift up to 2×PFOV due to the lens wheel jitter.

This last factor comes from three independent optical elements of ISOCAM, namely the lens wheel, the selection wheel and the filter wheel, which were mounted with a play to guarantee free movement (Vigroux et al. 1993; Cesarsky et al. 1996). As expected when this engineering solution was chosen, the different optical configurations can shift the position of the source on the detector by 2 pixels from the nominal pointing position.

We present the use of a statistical approach which increases the performance of cross-matching ISOCAM parallel sources with reference catalogues at optical or near-infrared wavelengths.

2. Comparison with reference catalogues

The USNO-A2.0 (Monet et al. 1998, 526 million sources), and the second incremental data release (IDR2, 162 million sources, covering 47% of the sky) of 2MASS (Skrutskie et al. 1997) were used as reference catalogues for astrometry.

As part of the preparation towards the final ISOCAM parallel catalogue, 24 000 images, containing 74 000 source candidates, were visually inspected to classify the source candidate via estimation of its pixel history as good, questionable or spurious (Ott 2002).

This classification was performed to derive cut-off parameters for a statistical cleaning of source candidates for the final catalogue, and the verification of processing strategy.

Send offprint requests to: S. Derriere,
e-mail: derriere@astro.u-strasbg.fr

[⋆] Based on observations with ISO, an ESA project with instruments funded by ESA Member States (especially the PI countries: France, Germany, The Netherlands and the UK) with the participation of ISAS and NASA.

Article published by EDP Sciences and available at http://www.aanda.org or http://dx.doi.org/10.1051/0004-6361:20030564

The main criterion to classify source candidates is their pixel history. In theory, the flux will increase after ISO finishes its slew, and the ISOCAM pixel starts to see the source at the begin of a pointing, and decrease when ISO slews away to a new pointing (see Fig. 1).

Additionally, within these 24 000 pointings, over 1300 visible *potential* source candidates, which were not detected by the automatic processing, were manually verified. The strongest undetected source found is a double star of flux 1.2 mJy and a S/N of 2, which validates the completeness of the source extraction.

As the statistical cleaning and merging of source candidates into unique sources is a significant data processing project in its own right, an easier-to-handle data-set was selected to derive the first cross-identifications: the data-set consists of all LW2 observations with 6″ PFOV, having at least eight readouts. Their pointing must be outside the galactic plane, with a galactic latitude outside ±20°. For tracking observations only the first pointing was included. Furthermore, all crowded regions (pointings showing extended structure or having more than six verified point sources) were excluded. The appropriateness of these heuristically chosen thresholds is confirmed by the first draft catalogue of statistically cleaned sources: over 98% of the sources classified as "good" by eye-balling fulfilled the statistical cleaning criteria (Schartel, priv. communication). More details will be provided in a forthcoming paper by Ott et al.

We ended with a sample of 7197 ISOCAM sources fulfilling the selection criteria, flagged as good candidates after eye-balling the pixel history.

For each ISOCAM source, its position on the detector in pixels coordinates (x, y) is known, together with the pointing parameters and the celestial coordinates (α, δ) (which derive directly from the nominal pointing position of the satellite). These celestial coordinates are affected by the three sources of error described in the introduction.

The main cause of astrometric inaccuracy being due to the instrument itself, we decided to study the difference in position between ISOCAM and reference sources in *detector pixel coordinates*, rather than in sky coordinates, to identify the consequences of the instrumental effects. Let us label i the ISOCAM sources. For each source i, detected at pixels (x_i, y_i), the following scheme was applied:

- all the n_i reference sources (optical or near infrared) within 60″ of position (α_i, δ_i) were searched;
- for each source j amongst n_i, the reference source sky coordinates (α_j, δ_j) were converted into ISOCAM pixel coordinates, using the pointing parameters of source i (not taking into account the ISOCAM distortion, but this has little effect on the result);
- the positional difference in pixels $(\Delta x_{ij} = x_i - x_j, \Delta y_{ij} = y_i - y_j)$ was computed.

Performing this operation on N ISOCAM sources, a total of $J = \sum_{i=1}^{N} n_i$ values are computed. We present in Fig. 2 the projection of all the $(\Delta x_{ij}, \Delta y_{ij})$, for the sample of 7197 ISOCAM sources, using the USNO as reference catalogue.

With a "perfect" instrument, one would expect a Gaussian distribution centered on (0,0). In our case, however, two strong peaks appear, centered on $(-1.25, +0.25)$ and $(+0.83, +0.25)$; they correspond to the rest positions, at the stops of the instrument lens wheel.

This clearly demonstrates that a nearest-neighbor approach for the cross-identification is not optimal in this case. We therefore decided to use a refined probabilistic approach to perform the cross-matching of ISOCAM sources with other catalogues.

3. The probability pattern

Figure 2 shows that counterparts will not be found preferably in the immediate vicinity of ISOCAM sources, but most likely they will lie at some distance corresponding to the shift induced by the lens wheel jitter. The probability of performing a good association will be at its highest in the two peaks of Fig. 2. Away from the peaks, a roughly uniform background is observed, as expected for a Poisson noise due to random spurious associations.

We transformed the distribution in Fig. 2 into a "probability map" $\mathcal{M}$. First, a square of side 6× PFOV in the $(\Delta x, \Delta y)$ plane was selected, and the 2D histogram was computed with a 1/6× PFOV bin size (corresponding to 1″ on the sky). The map is presented in Fig. 3.

The borders of the map were used to estimate the mean level of the background $\mathcal{B}$, and its variance σ. We finally obtain what we call the $n-\sigma$ probability pattern $\mathcal{P}$ by selecting all the bins fulfilling the condition $\mathcal{P}(x, y) > (\mathcal{M}(x, y) - \mathcal{B}) + n\sigma$. All the other bins in the pattern are set to zero and $\mathcal{P}$ is normalized to 1.

Figure 4 shows the 0σ and 2σ patterns. We have used these patterns as our selection criterion to perform the cross matching: rather than selecting the nearest (optical or near infrared) neighbor as the best possible counterpart of the ISOCAM source, we now select the reference source j with $(\Delta x_{ij}, \Delta y_{ij})$ corresponding to the highest (and non-zero) value in the pattern.

The non-zero area reduces to 138 and 72 square arcseconds, respectively, for the 0σ and 2σ patterns. This means, in the case of the 2σ pattern, that the useful search area is reduced by a factor of 18, compared to a blind search in the full square. This helps reduce the number of spurious associations, and increases the reliability of our cross-identifications.

4. Completeness and reliability

The quantification of the improvement due to the use of the probability pattern, as compared to the simple nearest-neighbor-based matching procedure, is difficult because one can never be completely sure whether an association made between two sources is correct or not.

Out of a sample of N ISOCAM sources, and for a given reference catalogue, only a fraction $f \cdot N$ has counterparts in the reference catalogue, because of the different sensitivities and wavelength ranges, the effect of proper motions, and the presence of solar system objects. In the ideal case, we would find the right associations for all the $f \cdot N$ sources, and identify the $(1-f) \cdot N$ sources not to be matched. In practice, the result of

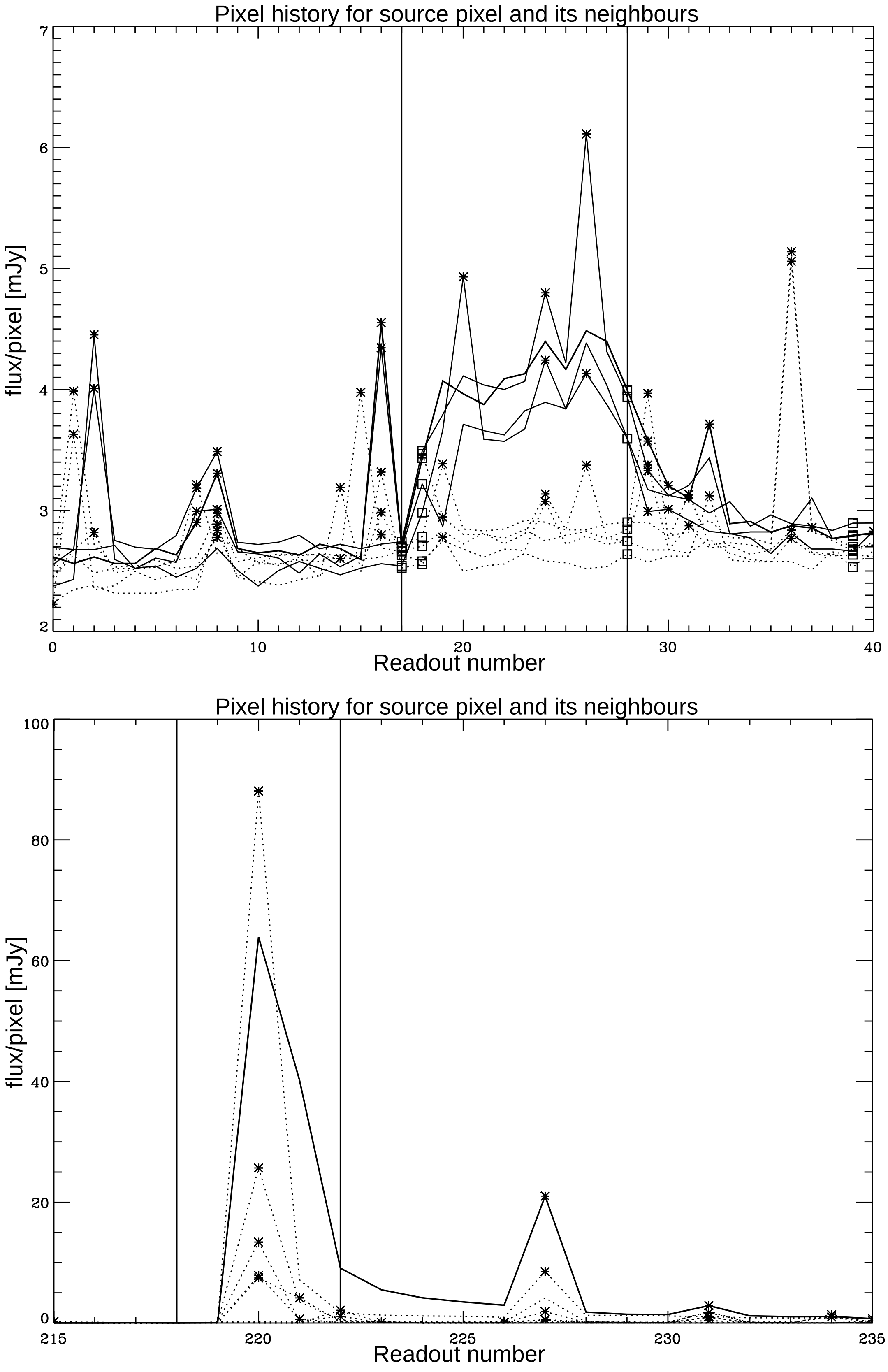

Fig. 1. Example of visual distinction between a bona-fide point source (top) and a spurious point-source (bottom) from its pixel history. The bold solid line corresponds to the strongest source pixel profile in both figures. In the top figure, other profiles correspond to the neighboring pixels receiving flux (solid lines), or not (dotted lines). In the bottom figure, the profiles corresponding to the 8 neighboring pixels are shown as dotted lines. Vertical lines indicate start of slew to new target. Slew periods when ISO did not acquire the target yet are marked by a box. Detected glitches are marked by a star.

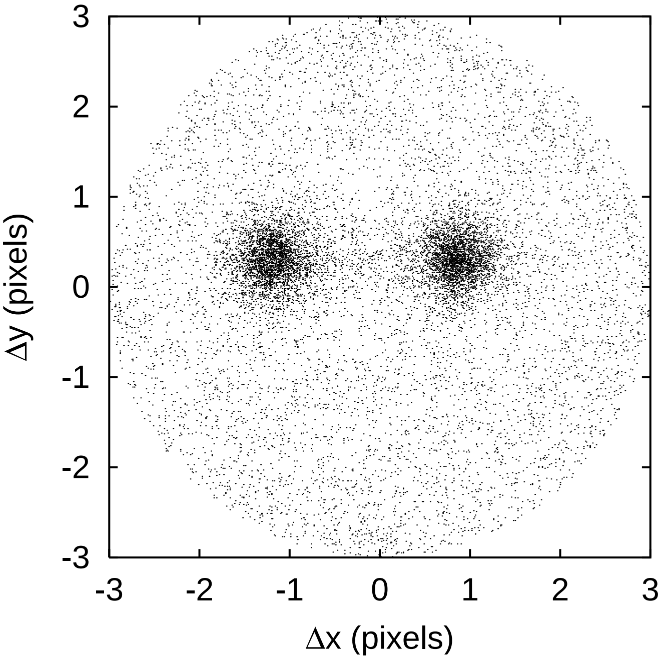

Fig. 2. Distribution of $(\Delta x_{ij}, \Delta y_{ij})$ for the 12 954 USNO matches within 18″ of the 7197 ISOCAM sources. The range is expressed in ISOCAM pixels (6″ width).

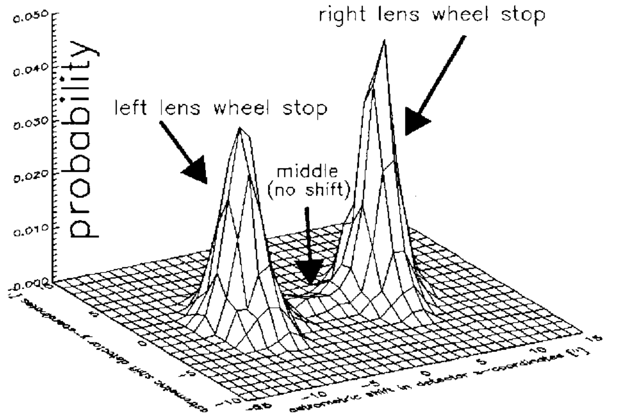

Fig. 3. Central part of the probability map $\mathcal{M}$ (normalized): 30 × 20 1″ bins in the $(\Delta x, \Delta y)$ plane.

the cross-identification is a number of associations A, selected as valid, and consisting of:

- R sources (among the fN) which are properly associated with their reference counterpart;
- W sources having a reference counterpart but wrongly associated;
- S sources among the $(1 - f)N$ sources without counterpart, randomly associated with unrelated reference sources.

The number of selected associations can be expressed as $A = R + W + S$.

When performing cross-identification, one attempts to find the best compromise between completeness and reliability. Completeness is defined as $c = R/fN$, the fraction of the fN sources having a counterpart that is properly associated. Reliability of the cross-matching is defined as $r = R/A$, the fraction of correct associations in the total number of associations (correct and incorrect).

Depending on the selection threshold we impose during the cross matching process, reliability and completeness will vary as illustrated in Fig. 5. The correlation between completeness and reliability depends on the decision criterion to distinguish matches from non-matches.

4.1. Statistical validation

Estimating the completeness and reliability of associations is not obvious, as the number of proper associations R is unknown. It is indeed impossible to identify every single proper association among the A selected associations. In addition, we do not know the value of f.

An alternative is to use a statistical model for associations: X_i associations having a probability p_i of being proper will contain on average $X_i \cdot p_i$ proper associations. If the p_i can be modeled, an estimate of the number of proper associations is given as $R = \sum_i X_i \cdot p_i$.

The likelihood of associations can be estimated using a Bayesian approach as shown by Sutherland & Saunders (1992). This method requires one to estimate priors for the probability distribution of magnitudes and distances of associated objects. This is usually done by comparison with background fields: the cross-identification of sources in the central field (same position on the sky) gives the sum of the desired probability distribution and some background noise due to random associations. The cross correlation of sources from the first catalogue with various fields of the reference catalogue (usually a few arcmin away) gives an estimate of the background noise.

However, as noted by Rutledge et al. (2000), this is only valid if the background fields have the same properties as the central field. In our study, this is not the case. ISOCAM sources correspond generally to bright optical or near-infrared objects, and the catalogues are biased around these objects. In USNO, there are often a large number of faint detections around bright objects (in the diffraction spike), while in 2MASS there is no source extraction around bright stars. This makes it impossible to estimate a reliable prior for the probability distribution of magnitudes in the reference catalogue.

Therefore, we decided to use only a distance criterion as a prior probability, and in addition, estimated the background contamination with geometric considerations in the central field itself, without using background fields.

The distance ρ between an association (located at $(\Delta x_{ij}, \Delta y_{ij})$ in the $(\Delta x, \Delta y)$ plane) and the nearest of the 2 peaks of Fig. 3 was selected as a key parameter to model p.

The histograms of the values of ρ were computed (with 0.5″ bin width, Fig. 6), for all possible associations, for the USNO and 2MASS catalogues. There are two components in these histograms: a peak at small values of ρ, corresponding to proper association, and a constantly rising component corresponding to random associations. This last component can be modeled. Given an homogeneous distribution of points with spatial density λ, the number of points at distance ρ from the

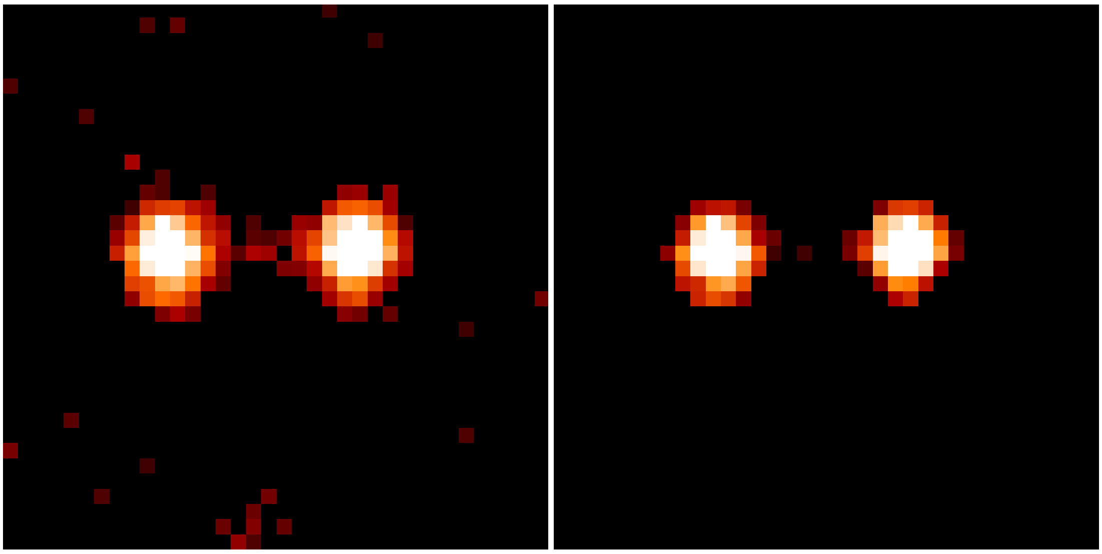

Fig. 4. 0σ (left) and 2σ (right) probability pattern $\mathcal{P}$ (36×36 1″ bins).

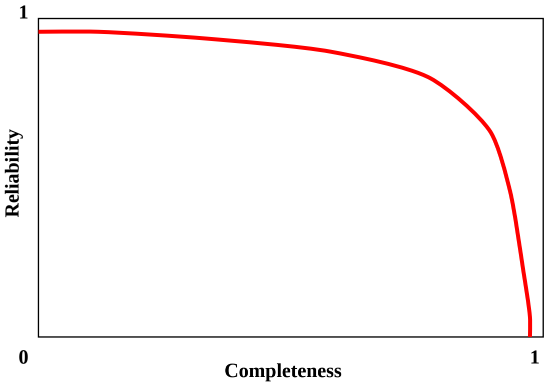

Fig. 5. Illustration of how completeness and reliability are correlated for a given method of cross matching, in the case where the spatial separation between counterparts is due to purely random errors.

closest of 2 reference points is

$$n(\rho) \cdot \mathrm{d}\rho = \begin{cases} 4\pi\lambda\rho \cdot \mathrm{d}\rho & (\rho < s) \\ 4\lambda\rho(\pi - \arccos(s/\rho)) \cdot \mathrm{d}\rho & (\rho > s), \end{cases}$$

the distance between the reference points being $2s$.

The density of sources in the reference catalogues is not constant. For each ISOCAM source i, the local density λ_i in the reference catalogues was computed (from the number of sources in 5′ radius around ISOCAM sources), and all the $n_i(\rho)$ were added to find the total expected random contribution (dashed line in Fig. 6, left panels).

The agreement with the observed distributions is very good. The probability of an association at distance ρ being proper was then simply derived by subtracting this random contribution from the observed total distribution. The resulting probabilities $p(\rho)$ are presented in the right panels of Fig. 6. The probability of a proper association with $\rho > 1$ pixel was clipped to zero. We modeled the probability with simple polynomials: $p(\rho) = 0.972 - 0.08\rho + 0.39\rho^2 - 2.24\rho^3 + 0.6\rho^4$ for USNO and $p(\rho) = 0.96 + 0.35\rho - 1.35\rho^2$ for 2MASS. We restricted the fit to second order for 2MASS in order to keep a smoothly decreasing function: higher order polynomials produced oscillations, because of the discrepancy at large distances. This discrepancy could be due to the incomplete coverage of the 2MASS IDR2 which introduces side effects not taken into account in our model.

The *FWHM* of the probability profile in Fig. 6 is close to 4″: this gives an estimate of the astrometric accuracy that will be achieved with the probability pattern.

The value of R will be estimated via $p(\rho)$: for all the selected associations j, ρ_j is computed, and $R = \sum_j p(\rho_j)$.

4.2. Nearest neighbor approach

The simplest and most widely used decision criterion for selecting cross-matching candidates is based on the search of nearest-neighbor (Bartlett & Egret 1998). The spatial distance d between sources is the only relevant parameter for decision in this case: the A associations which are closer than a threshold distance d_T are selected as matches, and all pairs with $d > d_\mathrm{T}$ are rejected as non-matches.

If there is no systematic offset between the positions of identical objects in the two catalogues, the reliability will be high for small values of d_T, but completeness will be low (upper left of Fig. 5). For increasing values of d_T, completeness will rise, but reliability will decrease, because of spurious associations that are more likely at larger distances.

We have searched for nearest neighbors of ISOCAM sources in the USNO and 2MASS catalogues. Figure 7 presents the results obtained with a 1′ search radius. The peak at short distances is mainly due to associations of the fN objects having a counterpart, while the tail at long distances is due to the $(1 - f)N$ objects without counterpart. For a local density λ_i in the reference catalogue, the density of nearest neighbor random matches at distance d derives from a 2D Poisson flux, and can be expressed as

$$\Phi(d, \lambda_i) = 2\pi\lambda_i d \exp(-\pi\lambda_i d^2).$$

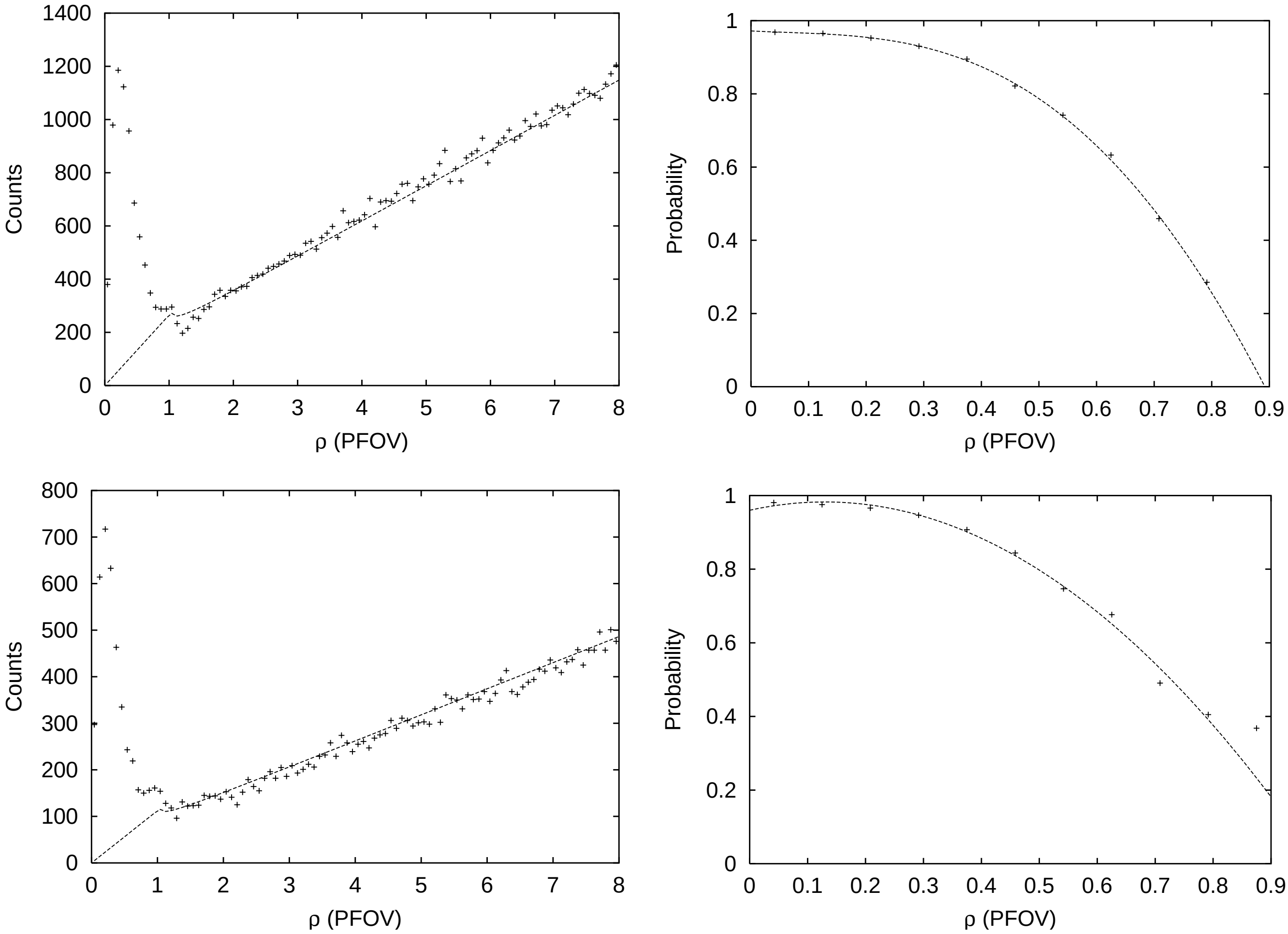

Fig. 6. Left: number of associations as a function of distance ρ to the nearest peak of Fig. 3 (0.5″ bin size). The dotted line corresponds to random associations. Right: derived probability $p(\rho)$ of an association being proper. Top row: USNO; bottom row: 2MASS.

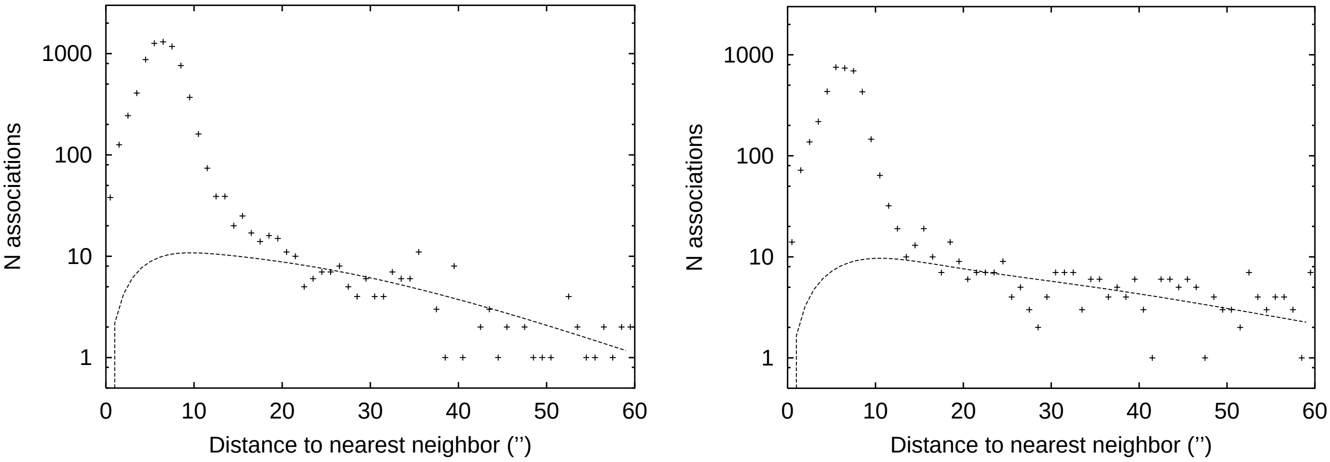

Fig. 7. Differential histogram (log scale) for the number of USNO (left) and 2MASS (right) matches as a function of distance to the nearest neighbor (bin width is 1″). The contribution of objects without matches is indicated by the dashed line.

Assuming that the fraction f of ISOCAM sources with a counterpart does not depend on the position on the sky (and thus does not depend on λ), one can express the total of all random matches as

$$T(d) = (1 - f) \sum_i N_i \cdot \Phi(d, \lambda_i).$$

Fitting this distribution on the tail of Fig. 7 gives an estimate of f. For USNO, the best fit is found for $f = 0.95$, indicating that 95% of ISOCAM sources have a counterpart in the USNO. For 2MASS, the fraction is $f = 0.91$, but this value might be affected by border effects, as the distribution of the IDR2 sources on the sky is very patchy, with many small areas not-covered.

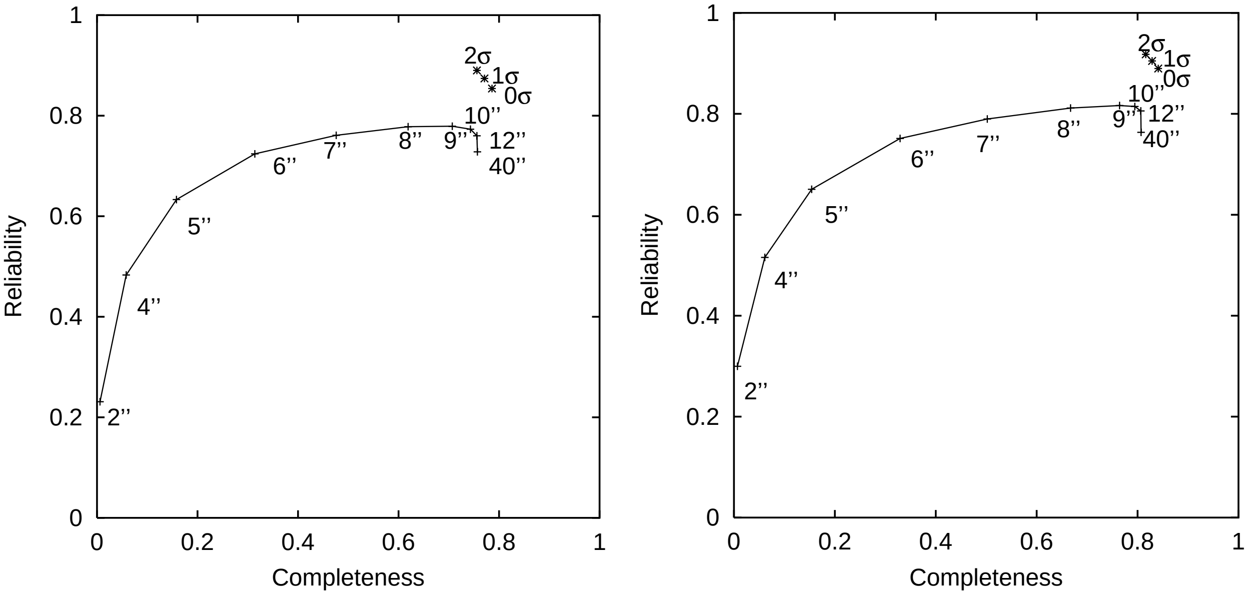

Fig. 8. Completeness versus reliability for nearest-neighbor (line) and probability pattern (stars) methods, for USNO (left) and 2MASS (right). The adopted thresholds are indicated for the different methods.

Table 1. Results of the nearest-neighbor method with $d_{\rm T} = 10''$: number of selected associations A, number of objects having a counterpart fN, estimated number of proper associations R, and corresponding completeness and reliability.

	USNO	2MASS
A	6570	3635
$f \cdot N$	6837	3726
R	5081	2961
c	0.74	0.79
r	0.77	0.81

Table 2. Results for the 1σ probability pattern.

	USNO	2MASS
A	6031	3413
$f \cdot N$	6837	3726
R	5273	3088
c	0.77	0.83
r	0.87	0.90

For different thresholds $d_{\rm T}$, we have selected all the $A_{d_{\rm T}}$ nearest neighbors in the reference catalogues. Distances to the nearest peak ρ_i were computed for every associations to estimate $R_{d_{\rm T}}$. A correction must be added for 2MASS, as the sky coverage is not complete: only 56.9% of the ISOCAM sources lie in an area covered by the 2MASS IDR2 (this value differs from the 47% sky coverage of the IDR2, because only ISOCAM sources outside the galactic plane were selected).

The estimated number of proper associations $R_{d_{\rm T}}$ made amongst the $A_{d_{\rm T}}$ selected association yielded an estimate of the changes of completeness and reliability of the nearest-neighbor method with $d_{\rm T}$. Results are presented in Fig. 8.

For very small distances, both c and r are low, because the associations are selected out of the 2 peaks, and are mainly spurious. Both c and r rise with $d_{\rm t}$, as the 2 peaks are progressively reached by the search radius.

The best compromise is around $d_{\rm T} = 10''$. The corresponding performances are reported in Table 1.

The optimal performance of the nearest neighbor are $c = 0.74$ and $r = 0.77$ for USNO, and $c = 0.79$ and $r = 0.81$ for 2MASS.

Further increasing $d_{\rm T}$ (up to $40''$, see Fig. 8) does not increase the number of proper associations R, so c saturates, and r decreases because of more spurious associations being selected.

4.3. Probability pattern

The probability pattern defined in Sect. 3 was used to perform the selection of associations between ISOCAM sources and the reference catalogues.

Among the possible j reference counterparts of ISOCAM source i, only the association corresponding the highest (and non zero) $\mathcal{P}(\Delta x_{ij}, \Delta y_{ij})$ was selected. Source i was not associated if all of the j had $\mathcal{P}(\Delta x_{ij}, \Delta y_{ij}) = 0$.

We have performed the selection of $A_{n\sigma}$ associations for different patterns (0, 1 and 2σ). Again, $R_{n\sigma}$ was estimated for each selection, and completeness and reliability have been computed (Fig. 8).

As expected, the most selective 2σ pattern gives the highest reliability, but is less complete, while the 0σ pattern gives the most complete sample, with a slightly less reliability.

Results for the 1σ pattern are given in Table 2: with $c = 0.77$ and $r = 0.87$ for USNO and $c = 0.83$ and $r = 0.90$ for 2MASS, results are much better than that achieved with the best selection of nearest neighbors.

5. Conclusions

The lens wheel jitter of the ISOCAM instrument gives an increased uncertainty on positions. This decreases the performance of cross-correlations of ISOCAM sources with reference catalogues.

For our sample of 7197 ISOCAM sources, the fraction having a real counterpart in the USNO and 2MASS catalogues was estimated to be 95% and 91%, respectively. This indicates that our ISOCAM source sample contains no more than 5% of spurious sources. Had the ISOCAM sample been contaminated by spurious detections, the fractions of sources having counterparts in reference catalogues would decrease. This would not qualitatively affect the results of the comparison between the two methods we used to perform associations.

Using a statistical approach, we proposed an original method to perform the associations. With the proposed probability pattern method to perform the cross identification with reference catalogues, the number of *selected* associations decreases, but the absolute number of *proper* associations increases.

As compared to what is achieved with a classical nearest-neighbor-based cross identification, the completeness for the selected associations increases from 74% to 77% for USNO and from 79% to 83% for 2MASS, while the reliability increases from 77% to 87% for USNO and from 81% to 90% for 2MASS.

This method could help reducing the astrometric uncertainty on the final ISOCAM parallel mode catalogue, and provide more reliable multi-wavelength information for the corresponding ISOCAM sources.

The method could be applied to all the cross-identification problems where the probability distribution is not symmetrical, to improve the results achieved by a classical nearest-neighbor approach.

This work also demonstrates some methods that will have to be used in the frame of the astronomical Virtual Observatory, to perform cross-identifications between pointed observations and large surveys, to derive multi-wavelengths catalogues.

Acknowledgements. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The ISOCAM data presented in this paper were analysed using 'CIA', a joint development by the ESA Astrophysics Division and the ISOCAM Consortium. The ISOCAM Consortium is led by the ISOCAM PI, C. Cesarsky.

References

Bartlett, J., & Egret, D. 1998, in New Horizons from Multi-Wavelength Sky Surveys (Baltimore - Maryland, 26-30 August, 1996), IAU Symp., 179, 437

Cesarsky, C. J., Abergel, A., Agnese, P., et al. 1996, A&A, 315, L32

Kessler, M. F., Steinz, J. A., Anderegg, M. E., et al. 1996, A&A, 315, L27

Monet, D., Bird, A., Canzian, B., et al. 1998, USNO-A V2.0, A Catalog of Astrometric Standards, Tech. rep., U.S. Naval Observatory, Washington DC

Ott, S. 2002, Ph.D. Thesis, Université Paris 6

Ott, S., Siebenmorgen, R., Schartel, N., et al. 2003, in preparation

Rutledge, R. E., Brunner, R. J., Prince, T. A., & Lonsdale, C. 2000, ApJS, 131, 335

Siebenmorgen, R., Abergel, A., Altieri, B., et al. 1996, A&A, 315, L169

Skrutskie, M. F., Schneider, S. E., Stiening, R., et al. 1997, in The Impact of Large Scale Near-IR Sky Surveys, April 1996, Tenerife (Spain), ed. F. Garzón, N. Epchtein, A. Omont, W. Burton, & P. Persi. (Kluwer), ASSL, 210, 25

Sutherland, W., & Saunders, W. 1992, MNRAS, 259, 413

Vigroux, L. G., Cesarsky, C. J., Boulade, O., et al. 1993, in Infrared Detectors and Instrumentation, ed. A. M. Fowler, Proc. SPIE, 1946, 281